Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Avaliação da plasticidade anatômica foliar de *Licania humilis* a partir de uma amostra pequena utilizando inferência bayesiana

Jéssica Carolina Zilio

Dissertação apresentada para obtenção do título de Mestra em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2023

Jéssica Carolina Zilio
Bacharel em Engenharia Agronômica

Avaliação da plasticidade anatômica foliar de *Licania humilis* a partir de uma amostra pequena utilizando inferência bayesiana

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Profa. Dra. **SÔNIA MARIA DE STEFANO PIEDADE**

Dissertação apresentada para obtenção do título de Mestra em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2023

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Zilio, Jéssica Carolina

Avaliação da plasticidade anatômica foliar de *Licania humilis* a partir de uma amostra pequena utilizando inferência bayesiana / Jéssica Carolina Zilio. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2023 .

59 p.

Dissertação (Mestrado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Cerrado 2. Adaptação 3. Estimação 4. Regressão . I. Título.

AGRADECIMENTOS

Agradeço a Deus pela vida e por nunca me desamparar.
Aos meus pais, Elaine e Rubens, por tudo o que fazem por mim.
Ao meu noivo Matheus, por tornar tudo ainda melhor.
Aos meus irmãos, Paola, Bianca e Miguel, pelo companheirismo.
À minha orientadora, Sônia, por ser uma profissional admirável e uma pessoa maravilhosa.
Ao meu amigo, Jhonathan, pelos ensinamentos que levo até os dias de hoje.
A todos os colegas com quem tive o prazer de conviver nesses anos.
A todos os professores, funcionários e colaboradores do Programa de Pós-Graduação em Estatística e Experimentação Agronômica da ESALQ.
O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

SUMÁRIO

Resumo 5
Abstract 6
1 Introdução 7
1.1 Objetivos 8
1.1.1 Objetivo geral 8
1.1.2 Objetivos específicos 9
2 Revisão Bibliográfica 11
2.1 A importância da amostragem na estatística 11
2.2 Desafios da amostragem na ecologia 12
2.3 Alternativas para o problema das pequenas amostras 13
2.4 Efeito do ambiente no desenvolvimento das espécies 14
2.5 Estimação de Parâmetros 16
2.5.1 Inferência clássica 16
2.5.2 Inferência Bayesiana 17
2.6 Métodos de Reamostragem 19
2.6.1 Teste de Aleatorização 19
2.6.2 Validação cruzada 20
2.6.3 *Jackknife* 20
2.6.4 *Bootstrap* 21
3 Material e Métodos 23
3.1 Estudo de simulação 23
3.1.1 A população inicial 23
3.1.2 Inferência clássica 23
3.1.3 Inferência bayesiana 24
3.2 Aplicação prática 25
3.2.1 Os dados reais 25
3.2.2 Análise 26
3.3 Regressão 26
4 Resultados 29
4.1 Estudo de simulação 29
4.2 Dados Reais 43
4.2.1 Análise exploratória 43
4.2.2 Análise frequentista 46
4.2.3 Análise bayesiana 48
4.3 Regressão 49
5 Conclusão 53
Referências 55

RESUMO

Avaliação da plasticidade anatômica foliar de *Licania humilis* a partir de uma amostra pequena utilizando inferência bayesiana

Plasticidade é a capacidade que um ser vivo apresenta de modificar suas estruturas de forma a adequar-se ao meio em que está inserido. O Cerrado brasileiro, por exemplo, é um bioma que apresenta muitas limitações para o desenvolvimento vegetal, sobretudo relacionadas à disponibilidade hídrica. Dessa forma, a literatura relata numerosas espécies nativas desse ambiente que apresentam plasticidade em uma ou mais estruturas. Este trabalho visa avaliar se a espécie *Licania humilis* apresenta plasticidade anatômica foliar quando observada em três diferentes ambientes: cerrado natural, cerrado em regeneração e sub-bosque de *Pinus*. Os atributos foliares avaliados para este propósito foram espessura da cutícula da face superior, espessura da parede celular periclinal externa, altura da célula epidérmica superior, espessura do mesofilo e altura da célula epidérmica inferior. Os dados reais disponíveis para estudo representavam amostras pequenas em relação ao que é recomendado para os testes convencionais, fazendo-se necessário utilizar métodos alternativos para este tipo de problema. Inicialmente, foi realizado um estudo de simulação comparando a estimação da média populacional utilizando inferência clássica e bayesiana para diferentes tamanhos de amostra e diferentes *prioris*. Em seguida, ambos os métodos foram novamente utilizados, desta vez partindo dos dados reais de *L. humilis* para os cinco atributos descritos. Finalmente, foi realizada uma regressão polinomial estimando valores da espessura do mesofilo com base na altura da célula epidérmica superior, simulando uma predição por variáveis correlacionadas. Para todas as formas de estimação de média utilizadas, a comparação entre os três ambientes foi realizada utilizando Análise da Variância e Teste de Tukey. Os resultados mostraram que a *L. humilis* apresenta plasticidade anatômica foliar quando observada nas diferentes regiões, respondendo adaptativamente às condições específicas de cada ambiente.

Palavras-chave: Cerrado, Adaptação, Estimação, Regressão

ABSTRACT

Evaluation of leaf anatomical plasticity of *Licania humilis* from a small sample using bayesian inference

Plasticity is the ability of an individual to modify its structures in order to adapt to the environment where it is inserted. The brazilian Cerrado biome, for example, presents many limitations to plant development, especially related to water availability. Therefore, the literature reports numerous Cerrado native species that present plasticity in at least one structure. This research aims to evaluate if the *Licania humilis* species presents leaf anatomical plasticity when occurring in three different environments: natural Cerrado, Cerrado in regeneration and *Pinus* understory. The leaf atributes evaluated for this purpose were the upper cuticle thickness, upper epidermal cell height, lower epidermal cell height, mesophyill height and upper outer periclinal wall. The real data available for this study consists in very small samples in relation to what is recommended for conventional tests, making it necessary to use alternative methods for this type of problem. Initially, it was performed a simulation study comparing the populational mean estimation using classical and bayesian inference to different sample sizes and different *priors*. Then, both methods were used again, this time starting from real *L. humilis* data for the five atributes mentioned. Finally, it was performed a polinomial regression estimating the mesophyill height values based on the upper epidermal cell height, simulating a prediction by correlated variables. For all the mean estimation methods used, the comparison between the three environments was performed using Analysis of Variance and Tukey's Test. The results showed that *L. humilis* presents leaf anatomical plasticity when observed in the three different regions, responding adaptivaly to the specific conditions of each environment.

Keywords: Brazilian Cerrado, Adaptation, Estimation, Regression

1 INTRODUÇÃO

Historicamente, a grande chave da vida é a adaptação. O que mantém um organismo vivo é um complexo conjunto de processos fisiológicos, físicos e químicos, cujo funcionamento é continuamente influenciado por condições externas. Ao longo dos anos, as espécies foram adquirindo, através de mutações e da seleção natural, novas características de forma a aumentar suas chances de sobrevivência no ambiente em que estão inseridas, processo conhecido como evolução.

Entende-se por sobrevivência a possibilidade de um indivíduo manter o funcionamento de seus processos vitais, tais como alimentar-se e reproduzir-se. Para os indivíduos do Reino Animal, a sobrevivência muitas vezes está relacionada à migração em busca de condições mais favoráveis, o que é possível uma vez que estes são dotados da capacidade locomotiva. Os indivíduos do Reino Vegetal, em contrapartida, são seres estáticos, não restando-lhes outra opção senão adaptar-se de forma a resistir às circunstâncias adversas (Marques *et al.*, 2000).

Nesse contexto, é chamada de plasticidade a capacidade que os seres vivos possuem de alterar suas estruturas a fim de tornar-se mais aptos ao ambiente em que se situam. Dessa forma, plantas de uma mesma espécie e até mesmo de material genético idêntico podem apresentar características diferentes de acordo com as condições externas às quais são expostas ao longo de seu desenvolvimento. Isso porque o fenótipo, que é a expressão individual de caracteres, é resultado da soma e interação dos fatores genotípicos e ambientais (Bradshaw, 1965).

O conceito de evolução fica mais claro quando tomamos como exemplo o bioma do Cerrado. Embora seja caracterizado pelo seu clima quente e seco, com altos níveis de insolação, solos arenosos, ácidos, nutricionalmente pobres, com baixa capacidade de retenção de água e alta concentração de alumínio, o Cerrado brasileiro é considerado o que possui maior riqueza em biodiversidade entre as savanas do mundo todo (Klink and Machado, 2005). Isso se deve ao fato de que as espécies nativas desse bioma são resultado de um longo processo evolutivo, conferindo-lhes características de plantas xeromórficas para que pudessem se adaptar às condições hostis a serem enfrentadas.

Dentre essas características, é possível observar um sistema radicular muito bem desenvolvido, possibilitando que a planta consiga acessar água em grandes profundidades, além de acumular reservas energéticas para uma eventual necessidade de rebrotamento. Em relação às plantas lenhosas, o seu caule é consistente e recoberto por uma camada de súber, protegendo a planta contra queimadas, que são bastante recorrentes nessse bioma. Além disso, as espécies nativas do Cerrado possuem folhas recobertas por uma espessa cutícula, além de concentrar seus estômatos na face abaxial, diminuindo substancialmente a perda de água pela transpiração (Bieras and Sajo, 2009).

O conceito de plasticidade, por sua vez, esclarece o fato de que uma planta típica do Cerrado pode, quando observada em um ambiente diferente daquele para o qual ela é adaptada, modificar algumas de suas estruturas conforme a necessidade e conveniência. É comum que, em um plantio em larga escala de espécies exóticas para fins comerciais, como *Pinus* ou eucalipto, sejam encontradas unidades remanescentes de espécies nativas desenvolvendo-se em seu sub-bosque, oriundas do brotamento das gemas presentes no sistema subterrâneo. As condições encontradas por esses indivíduos, tais como sombreamento acentuado e alta competição interespecífica, são muito distintas daquelas para as quais suas características foram moldadas, fazendo-se necessária uma nova forma de adaptação (Via and Lande, 1985).

A crescente e intensa exploração do Cerrado para fins comerciais é um fator que tem causado um comprometimento da flora nativa, levando esta a se adaptar para garantir sua sobrevivência, de forma

que já foi constatada cientificamente plasticidade em várias estruturas, sobretudo nas folhas (Fank-de Carvalho *et al.*, 2010), para um grande número de espécies provenientes desse bioma. Visto que esse tipo de interferência humana tem causado graves consequências para o ecossistema como um todo, algumas medidas já vem sendo tomadas na esperança de reverter essa infeliz situação.

Um exemplo prático ocorreu na Estação Ecológica de Santa Bárbara (EEcSB), situada na cidade de Águas de Santa Bárbara, no estado de São Paulo. Nessa área, que naturalmente era abrangida pelo bioma do Cerrado, foi implantado o cultivo de *Pinus* em meados dos anos 1970. Durante o desenvolvimento dessas plantas, alguns indivíduos de espécies nativas do Cerrado conseguiram sobreviver e desenvolver-se em um extrato abaixo do dossel de *Pinus*, doravante tratado como sub-bosque. Anos mais tarde, parte dessa área teve sua plantação removida através de corte raso e queimada superficial dos vestígios dessas árvores, permitindo o rebrotamento das espécies nativas em uma tentativa de restabelecer a vegetação original da região.

Partindo-se do conhecimento do fenômeno da plasticidade vegetal, é plausível formular uma hipótese de que os indivíduos que vivem no sub-bosque de *Pinus* apresentam características diferentes daqueles que rebrotaram após o seu corte, bem como daqueles que desenvolveram-se no Cerrado natural. Para que essa hipótese seja avaliada efetivamente, é necessário identificar evidências baseadas na análise estatística dos dados observados, com a finalidade de se chegar a uma melhor compreensão do real comportamento dessas espécies nas condições do estudo, bem como buscar uma explicação biológica para o fenômeno observado (Rodrigues *et al.*, 2007).

Coletar as informações em quantidade e qualidade necessárias para que a análise seja realizada de forma satisfatória é, contudo, uma das etapas mais desafiadoras dos estudos na área de ecologia. Embora seja de interesse do pesquisador obter o maior volume de dados possível, existem muitas limitações na prática que impedem a coleta de grandes amostras (Van de Schoot and Miocević, 2020). Um desses casos é quando o método de mensuração da variável de interesse é destrutivo, como para determinados atributos do sistema radicular, por exemplo, que implicam muitas vezes em sacrificar a planta por inteiro. No circunstância de um ensaio planejado para tal finalidade, a destruição do material não representa impedimento algum, todavia quando o objeto de interesse são plantas em seu habitat natural, os problemas vão muito além de questões éticas e ambientais (Butler, 2015).

Existem também casos em que o grupo de interesse é naturalmente pequeno, como no exemplo de indivíduos nativos do Cerrado sobreviventes vivendo no sub-bosque de um cultivo comercial. Outra situação é quando o material de interesse é abundante, porém, sua mensuração é complexa ou financeiramente inviável, devido à demanda de insumos de alto custo, mão-de-obra especializada, dentre outros fatores. Isso tudo implica na necessidade de se buscar métodos alternativos para que seja possível realizar uma interpretação correta e confiável dos dados, seja através de novos métodos de amostragem, diferentes abordagens da estatística inferencial, dentre outras possibilidades.

1.1 Objetivos

1.1.1 Objetivo geral

O objetivo deste trabalho é verificar o efeito que três diferentes ambientes (Cerrado natural, Cerrado em regeneração e sub-bosque de *Pinus*) exercem sobre a anatomia e micromorfologia foliar de uma espécie vegetal nativa do Cerrado a partir de uma pequena amostra, utilizando as abordagens frequentista e bayesiana para estimação dos parâmetros de interesse.

1.1.2 Objetivos específicos

- Avaliar o desempenho do método bayesiano quando comparado ao método clássico de estimação de parâmetros através de um estudo de simulação, bem como a análise de dados reais da espécie *Licania humilis*;
- Avaliar a utilização de uma variável na predição de outra utilizando regressão polinomial e como isso implica no resultado final do teste de comparação de médias;
- Verificar, baseado nos resultados das análises estatísticas mencionadas, se a espécie *L. humilis* apresenta plasticidade anatômica foliar para cinco atributos: espessura da cutícula da face superior, espessura da parede celular periclinal externa, altura da célula epidérmica superior, espessura do mesofilo e altura da célula epidérmica inferior.

2 REVISÃO BIBLIOGRÁFICA

2.1 A importância da amostragem na estatística

Um estudo científico, nas mais diversas áreas do conhecimento, surge com a necessidade do pesquisador em responder determinada pergunta a respeito de seu objeto de estudo. Dessa forma, o Método Científico existe com o objetivo de padronizar essa busca por respostas, sendo um conjunto de etapas e procedimentos para se chegar sistematicamente ao conhecimento científico com segurança e credibilidade (Rodrigues *et al.*, 2007).

O Método Científico, de acordo com Diggle *et al.* (2011), baseia-se na teoria e observação como pilares principais. Sucintamente, o processo tem início quando o pesquisador levanta uma hipótese acerca de seu objeto de estudo, a fim de responder uma ou mais perguntas a ele relacionadas. Em seguida, tem-se início a etapa de observação, na qual os dados são coletados através da realização de experimentos ou pesquisas observacionais, dependendo dos objetivos do estudo. Obtidos os dados, estes são analisados estatisticamente para se testar a hipótese inicial, oferecendo ao pesquisador evidências para que este finalmente desenvolva uma teoria a respeito de sua investigação.

Desse modo, a Estatística é um instrumento fundamental para a realização do Método Científico, pois se faz presente nas etapas de observação (escolha de métodos de amostragem apropriados) e de análise dos dados (inferência estatística) (Reis *et al.*, 1999). Como foi apontado em Diggle *et al.* (2011), a Estatística não tem como finalidade provar teorias, mas oferecer técnicas eficientes para se testar a consistência de hipóteses, trazendo evidências para sustentar um argumento com um grau maior ou menor de incerteza.

O escopo de uma investigação científica é, portanto, obter determinado conhecimento a respeito de uma população ou fenômeno, sendo geralmente representado por parâmetros. O fato é que nunca é possível observar a população em sua totalidade, fazendo-se necessário lançar mão de técnicas de amostragem (Van de Schoot and Miocević, 2020). Para definir em poucas palavras, amostras são subconjuntos extraídos de uma população com a finalidade de serem analisados, enquanto que a amostragem é o método através do qual as amostras são obtidas (Acharya *et al.*, 2013). A ideia é que seja possível tirar conclusões inicialmente a respeito da amostra e, por meio da inferência estatística, projetar esse conhecimento de forma que se torne uma generalização que atenda toda a população (Thompson, 2012).

Para que essa generalização seja possível, é imprescindível garantir que a amostra preserve as principais características da população que lhe deu origem, ou seja, que seja tão representativa quanto possível. Em geral, quanto mais elementos uma amostra dispor, maior será o seu poder estatístico e, consequentemente, melhor será o resultado final, embora esse não seja o único fator determinante para a qualidade da análise (Prajapati *et al.*, 2010). Além disso, alguns modelos de alta complexidade demandam um grande volume de dados para que suas pressuposições sejam atendidas, caso contrário não é possível assegurar a sua prestabilidade.

O Teorema do Limite Central é um dos mais importantes da inferência estatística, pois dá base a uma série de testes paramétricos que são amplamente utilizados na estatística moderna. Segundo esse teorema, as médias amostrais seguem uma distribuição aproximadamente normal em torno do valor da média populacional, independente de qual seja a distribuição de probabilidade da população. Conforme o tamanho amostral cresce, mais a distribuição das médias amostrais converge para uma distribuição normal e maior é a confiança em torno da média populacional (Kwak and Kim, 2017). Essa convergência, no

entanto, apenas pode ser garantida quando o tamanho das amostras for suficientemente grande (Dantas, 2013).

Com relação à coleta dos dados, existem dois tipos de estudo dos quais é possível obtê-los: o experimento manipulativo ou ensaio, que é planejado e instalado com o objetivo de testar determinada hipótese, sendo todos os fatores externos devidamente controlados; e o estudo observacional, em que o pesquisador não interfere nem controla os fatores, apenas observa a população e coleta os dados utilizando o método de amostragem mais conveniente (Rosenbaum, 2005).

Em geral, é preferível realizar experimentos manipulativos, pois neles é possível dissociar precisamente os efeitos dos fatores em estudo, uma vez que todas as condições ambientais são controladas de forma a minimizar o que se entende por variação do acaso. Todavia, sabe-se que muitos processos fisiológicos são influenciados por um conjunto de fatores que não podem ser desvinculados, já que a natureza é um complexo impossível de se reproduzir em casas de vegetação, fazendo-se então necessário realizar a coleta dos dados através da observação das plantas em seu habitat natural (Hairston, 1989).

Contudo, é comum que a coleta de grandes amostras não seja possível. Na prática, são numerosos os fatores que podem limitar uma amostragem eficiente, tais como grupos de interesse naturalmente pequenos ou dispersos, métodos de coleta complexos, destrutivos ou que exigem um alto custo operacional, dentre outros (Van de Schoot and Miocević, 2020). Nesses casos, é necessário que o pesquisador considere formas alternativas de testar as hipóteses de seu estudo, sem que para isso sofra com a perda de qualidade dos resultados.

2.2 Desafios da amostragem na ecologia

Nas ciências florestais, é cotidiano deparar-se com espécies arbóreas de ciclo longo, o que representa mais uma adversidade para a experimentação. Em muitos casos, é inviável conduzir um experimento por vários anos, fazendo-se necessário utilizar plantas já desenvolvidas, que todavia não foram plantadas seguindo delineamentos ou sequer conduzidas de forma a controlar os fatores externos. Além disso, plantios de espécies florestais demandam áreas extensas, devido sobretudo ao porte das árvores e ao espaçamento necessário entre elas (Scheiner and Gurevitch, 2001).

Nas ciências agrárias, por sua vez, a realização de experimentos é bastante frequente, já que a maioria das espécies agrícolas apresentam ciclo curto, além de não haver requisição de áreas tão extensas (Poehlman, 2013). No entanto, as questões de dificuldade operacional e do custo dos materiais podem configurar em barreiras que limitam uma coleta ampla de dados.

Nas ciências ecológicas, a complexidade se intensifica ainda mais, na medida em que experimentos manipulativos, por definição, não representam bem o ambiente natural (Hairston, 1989). A maior parte das pesquisas na área de ecologia são, portanto, estudos observacionais, afinal, compreender uma espécie implica em observar seu comportamento inerente ao seu habitat, onde interage com um sistema riquíssimo que exerce um papel fundamental no desenvolvimento da espécie (Resetarits and Bernardo, 2001). Processos como competição interespecífica, intraespecífica, predação e mutualismo são as principais formas de interação entre seres vivos, e interferem ativamente na dinâmica de populações (Gotelli, 2009).

Dados de pesquisas na área de ecologia, sejam eles provenientes de estudos observacionais ou de ensaios em laboratório, estufa ou no campo, compartilham da característica de apresentar diversas limitações. Dentre outras razões, delineamentos altamente desbalanceados, indisponibilidade de áreas

grandes, escassez de recursos e de tempo acessível são fatores que contribuem para que, nesse tipo de estudo, muitas pressuposições básicas da estatística sejam seriamente violadas (Scheiner and Gurevitch, 2001).

Além disso, existem algumas características de interesse científico cujos métodos de avaliação são destrutivos, tais como avaliação das raízes, cálculo de biomassa da parte aérea, dentre outros (Shepherd, 2017). Essa questão envolve sérios problemas éticos, pois muitas vezes implica em desmatamento de áreas de preservação, perturbação de ecossistemas e uma série de outros princípios cuja infração não pode ser justificada ou amparada pela necessidade em se fazer ciência (Butler, 2015).

Com os avanços nas áreas de pesquisa e de computação, a análise de dados ecológicos também tem passado por transformações positivas, embora o problema das pequenas amostra perdure até os dias atuais (Fox *et al.*, 2015). Na ciência, cada pesquisa possui suas próprias singularidades a serem consideradas, de modo que não existe um único procedimento eficiente em todos os casos. A literatura aborda alguns métodos constantemente utilizados para se contornar o problema de escassez de dados, com enfoque em diversas áreas, dentre elas a ecologia.

2.3 Alternativas para o problema das pequenas amostras

Uma possível saída para o caso de grupos de interesse naturalmente pequenos ou dispersos é escolher modelos de menor complexidade e que não tenham como premissa o Teorema do Limite Central, que possam ser ajustados apropriadamente com um volume inferior de dados. Todavia, essa estratégia pode comprometer a análise, na medida em que o modelo simplificado pode não ser o que melhor representa o comportamento do fenômeno de estudo. Esse cenário frequentemente não é favorável para a realização dos testes de hipóteses, tampouco para responder satisfatoriamente às perguntas do pesquisador (Van de Schoot and Miocević, 2020).

Ainda tratando-se de escassez de dados, para contornar as limitações associadas a casos em que as premissas do modelo não são atendidas, existe a possibilidade de se estimar o mesmo parâmetro repetidas vezes partindo de subconjuntos de uma mesma amostra, utilizando para isso o processo de reamostragem (James *et al.*, 2013). Dessa forma, determina-se a precisão de estimadores, além de identificar a ocorrência indesejada de sobreajuste, também conhecido como *overfitting*. Como mencionado por Yu (2002), a literatura trata de diversos métodos de reamostragem, sendo os quatro principais o Teste de Aleatorização, Validação Cruzada, Jackknife e Bootstrap.

Há casos em que a limitação da amostragem não se dá pelo tamanho da população original, mas pela dificuldade em se coletar os dados. Uma alternativa para quando a variável de interesse é de difícil mensuração ou requer métodos destrutivos é utilizar variáveis auxiliares fortemente correlacionadas com o fator de interesse e cuja mensuração seja consideravelmente mais simples. Se essa prática for viável no contexto do estudo, é possível obter amostras grandes o suficiente para que possam ser utilizados métodos estatísticos convencionais nas análises (Liang and Zeger, 1993).

Essa estratégia se faz presente, por exemplo, na amostragem por conjuntos ordenados, em que é possível ranquear os elementos de uma amostra segundo algum critério eficiente referente à variável correlacionada e, assim, estimar a variável de interesse (Taconeli and Giolo, 2020). Outra alternativa é ajustar uma regressão polinomial entre duas variáveis fortemente correlacionadas, de forma que seja possível predizer a variável de interesse utilizando apenas as observações da variável preditora (Ostertagová, 2012).

Para concluir, independente de qual seja a razão das amostras diminutas, é possível valer-se da abordagem bayesiana para realizar a inferência e estimar o parâmetro de interesse com eficácia. Isso porque a disponibilidade de *prioris* informativas na inferência bayesiana acaba por dispensar um grande volume de dados, uma vez que a influência que a amostra exerce sobre a *posteriori* é proporcional ao seu tamanho (Van de Schoot and Miocević, 2020).

Adicionalmente, vale ressaltar que o que garante de fato o êxito da análise é, de forma primordial, a representatividade da amostra. Logo, as metodologias apresentadas devem ser tidas como último recurso, quando realmente não for possível coletar amostras de tamanho adequado. Em outras circunstâncias, não é recomendado abdicar inadvertidamente de uma amostragem adequada apenas com a finalidade de utilizar métodos alternativos, sob risco de comprometer os resultados.

2.4 Efeito do ambiente no desenvolvimento das espécies

Todo vegetal vivo é microscopicamente formado por células que contêm em seu núcleo material genético responsável por definir suas características. No entanto, indivíduos com material genético idêntico podem apresentar-se distintos em vários aspectos durante seu desenvolvimento. A isso atribui-se o fato de que a espressão genética varia de acordo com o ambiente em que o indivíduo está inserido, sobretudo em se tratando de características quantitativas ou poligênicas, as quais são altamente influenciadas por fatores externos (Mackay *et al.*, 2009). Dessa forma, o fenótipo é resultado da interação entre genótipo e ambiente (Griffiths *et al.*, 2010).

O que determina o quanto a expressão fenotípica pode variar, morfológica ou fisiologicamente, de acordo com o ambiente com o qual o genótipo interage é denominado de plasticidade (Bradshaw, 1965). Esta apresenta-se de formas distintas para diferentes espécies e seus atributos, podendo um indivíduo de uma espécie apresentar alta plasticidade para determinada característica ao ser exposto em diferentes ambientes, e baixa ou nenhuma plasticidade para outra característica. As estruturas vegetais mais passíveis a alterações devido a adversidades ambientais são as folhas (Fank-de Carvalho *et al.*, 2010).

Essa relevante propriedade dos vegetais pode ser interpretada como um mecanismo adaptativo, o qual confere ao indivíduo a capacidade de sobrevivência em ambientes tidos como hostis para o seu desenvolvimento (Via and Lande, 1985). Dentre os fatores ambientais que estimulam a plasticidade em plantas, destacam-se disponibilidade hídrica e de nutrientes no solo (Pompelli *et al.*, 2019), condições edafoclimáticas e de luminosidade (Rossatto and Kolb, 2010), além da proteção contra doenças e predadores (Rôças *et al.*, 1997).

O Cerrado brasileiro, embora seja um dos biomas mais ricos em biodiversidade do mundo (Klink and Machado, 2005), é caracterizado pela restrição hídrica periódica, solos nutricionalmente pobres e ocorrência de queimadas (da Silva *et al.*, 2008). Desse modo, as plantas nativas desse bioma apresentam uma série de traços xeromórficos (Bieras and Sajo, 2009), tais como sistema radicular profundo e desenvolvido, caules espessos e tortuosos revestidos de uma camada suberosa, estruturas foliares adaptadas, dentre outros. Isso tudo porque, evolutivamente, essas espécies tiveram que adaptar-se a um ambiente com severas limitações (Marques *et al.*, 2000).

Em grande parte das espécies vegetais, a folha é o único e principal órgão responsável pela produção de energia através da realização de fotossíntese, e seu metabolismo perfeito está associado a diversos fatores internos e externos à planta. Toda folha é constituída por algumas estruturas básicas, cada

qual exercendo sua função no funcionamento do organismo de que faz parte. Diferentes espécies vegetais podem apresentar-se morfologicamente distintas em vários aspectos, entretanto as plantas dicotiledôneas compartilham de algumas estruturas básicas, como as representadas na Figura 2.1.

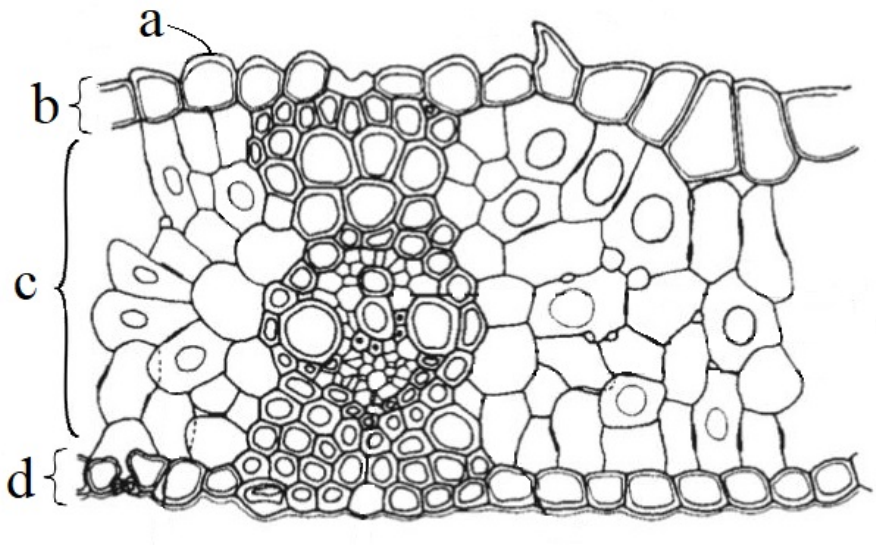

Figura 2.1. Estrutura interna de uma folha

(a) Cutícula foliar; (b) Epiderme superior; (c) Mesofilo; (d) Epiderme inferior. Figura adaptada de Cutler *et al.* (2009)

Dentre os papeis exercidos por essas estruturas na planta, vários encontram-se fortemente associados à resiliência vegetal em relação às adversidades ambientais (Hanson, 1917). A epiderme foliar, por exemplo, é o tecido que envolve toda a superfície adaxial e abaxial da folha, cumprindo tanto o papel de revestimento como também funções fisiológicas mais complexas (Appezzato-da Glória and Carmello-Guerreiro, 1992). Isso porque a camada serve como barreira física contra o ataque de patógenos e choques mecânicos, além de possuir estruturas especializadas em realizar trocas gasosas promovendo o balanço hídrico, chamados estômatos (Javelle *et al.*, 2011).

Somado a isso, a epiderme é recoberta pela cutícula, uma fina camada de compostos lipídicos, sobretudo a cutina, cujas propriedades hidrofóbicas a permitem exercer sua principal função, que consiste em proteger a folha contra a perda excessiva de água (Ristic and Jenks, 2002). Além disso, essa estrutura também auxilia na reflexão do excesso de luz e na contenção do ataque de fungos e da herbivoria (Fank-de Carvalho *et al.*, 2010).

Localizado entre as epidermes superior e inferior, o mesofilo é constituído sobretudo pelos parênquimas clorofilianos, sendo o grande responsável pelo processo de fotossíntese nas folhas (Cutler *et al.*, 2009). Estrategicamente, essa estrutura pode apresentar variações entre as espécies no que diz respeito ao volume das células, área superficial celular, proporção entre o parênquima paliçádico e lacunoso, dentre outros. O que determina as principais características do mesofilo é o ambiente para o qual a espécie é adaptada, principalmente em relação à incidência luminosa (Ivanova and P'yankov, 2002) e à disponibilidade hídrica (Pompelli *et al.*, 2019).

2.5 Estimação de Parâmetros

Para melhor compreender o comportamento de determinada característica de uma população pela qual o pesquisador tem interesse, é comum retratá-la por meio de um modelo, ou seja, uma representação matemática que visa descrever um fenômeno, seja ele natural ou artificial, através de expressões numéricas. Um modelo, por sua vez, pode ser determinístico ou não-determinístico, este último também conhecido como estocástico (Rohatgi, 2013).

Um modelo determinístico é caracterizado pela dependência exclusiva que o resultado apresenta dos fatores que o produzem. Dessa forma, ao observar o mesmo fenômeno repetidas vezes, sempre o mesmo resultado é contemplado. Em contrapartida, um modelo não-determinístico ou estocástico incorpora a chamada variação do acaso (ou erro experimental), na medida em que repetidas observações do mesmo fenômeno produzem valores diferentes da variável resposta, de forma que não é possível predizê-la com exatidão (Silvey, 1975).

Em situações aplicadas no âmbito agronômico e ecológico, a característica de interesse frequentemente apresenta comportamento estocástico, sendo muito comum representá-la através de uma variável aleatória X, a qual tem seu comportamento dado por meio de uma função de densidade de probabilidade $f(x/\boldsymbol{\theta})$ moldada por um ou mais parâmetros $\boldsymbol{\theta}$ (Magalhães, 2006).

A finalidade da inferência estatística é, portanto, encontrar e avaliar métodos para estimar os parâmetros de interesse da população através de informações obtidas por meio de amostras (Casella and Berger, 2021). Todavia, não existe um único procedimento eficiente para se obter boas estimativas, de forma que é possível seguir diferentes linhas e encontrar resultados equivalentes. Neste trabalho será tratado das abordagens clássica e bayesiana, melhor explicadas a seguir.

2.5.1 Inferência clássica

A inferência clássica ou frequentista é um ramo da estatística inferencial que tem como principal objetivo obter conclusões acerca de um fenômeno com base na frequência com que um evento ocorre em dados amostrais. Apoia-se na ideia de que probabilidade e frequência andam lado a lado, uma vez que o fenômeno de interesse deve manifestar-se na amostra da mesma forma que se manifesta na população que a originou (Wagenmakers *et al.*, 2008).

Na abordagem frequentista, um estimador pontual ($\hat{\theta}$) consiste basicamente em uma estatística, isto é, uma função da amostra aleatória que seja independente de outros parâmetros desconhecidos (Casella and Berger, 2021). Para que um estimador seja considerado apropriado, ele deve possuir algumas propriedades, sendo: suficiente (quando a distribuição condicional da amostra dado a estatística não depende de θ), não viciado (que não é tendencioso, ou seja, a esperança da estatística é igual a θ), consistente (quando para tamanho amostral tendendo ao infinito, converge em probabilidade para θ e em variância para zero) e eficiente (quando a variância da estatística é igual ao limite inferior da cota de Cramer-Rao) (Bolfarine and Sandoval, 2001).

A estimativa, por sua vez, é o valor numérico calculado fazendo uso do estimador a partir do conjunto de dados amostrais, com o intento de ser o mais próximo possível do valor real do parâmetro de interesse (Casella and Berger, 2021). Atendendo à propriedade da consistência, à medida que o número de elementos da amostra aumenta, a estimativa vai convergindo e aproximando-se cada vez mais do valor verdadeiro do parâmetro, de tal sorte que se a amostra for suficientemente grande, a obtenção de boas estimativas é totalmente possível (Prajapati *et al.*, 2010).

Além disso, algumas propriedades dos estimadores somente podem ser verificadas diante de amostras com muitos elementos, visto que são relacionadas ao comportamento assintótico do estimador, ou seja, ao cálculo dos limites para quando o número de elementos da amostra tende ao infinito (Bolfarine and Sandoval, 2001). O mais famoso exemplo é o Teorema do Limite Central que, como já foi explanado, somente é válido quando o tamanho amostral for grande o suficiente para que a convergência seja alcançada. À vista disso, é importante considerar que muitas vezes amostras de tamanho limitado ferem os princípios da estatística frequentista e podem, assim, inviabilizar todo o processo seguinte de avaliação de hipóteses.

Uma das técnicas mais utilizadas para a estimação de parâmetros na inferência clássica é o Método da Máxima Verossimilhança. O procedimento consiste em construir a função de verossimilhança por meio do produtório da distribuição condicional de cada elemento da amostra dado θ, aplicar o logaritmo e encontrar o ponto cuja primeira derivada iguala-se a zero e a segunda derivada é negativa, caracterizando o ponto de máximo da função (Cam, 1990). O valor encontrado será a melhor estimativa para o parâmetro, pois maximiza a probabilidade de que a amostra ocorra a partir dos dados em análise. Em contrapartida, o método costuma ser bastante sensível a mudanças na amostra, de modo a resultar em diferentes estimativas conforme variações nos dados (Casella and Berger, 2021).

Além do método da Máxima Verossimilhança, existem outros amplamente difundidos com a finalidade de se obter estimadores, tais como o Método dos Momentos e o Método dos Mínimos Quadrados (Bolfarine and Sandoval, 2001). A escolha da metodologia mais apropriada nem sempre é simples, pois frequentemente consiste na obtenção e avaliação de mais de um $\hat{\theta}$. Na busca pelo melhor estimador não viesado, geralmente opta-se por aquele dentre os não tendenciosos com a menor variância (Casella and Berger, 2021).

2.5.2 Inferência Bayesiana

A inferência bayesiana é um ramo da estatística inferencial fundamentada no Teorema de Bayes, e tem sido cada vez mais utilizada em diversas áreas da ciência, inclusive na ecologia (Banner *et al.*, 2020). Partilhando do mesmo princípio do aprendizado de máquinas (ou *machine learning*, como vem se popularizando), em que uma máquina vai aperfeiçoando paulatinamente seu funcionamento à medida que vai aprendendo com novas informações, a inferência bayesiana utiliza informações a *priori* e as atualiza com base em novas amostras, obtendo assim a *posteriori*. Esta, por sua vez, assumirá o papel de *priori* assim que se tenha acesso a novos dados, estabelecendo assim uma relação cíclica (Van de Schoot and Miocević, 2020).

$$p(\theta/dados) = \frac{p(dados/\theta)p(\theta)}{p(dados)} \tag{2.1}$$

O Teorema de Bayes, baseado nas leis da probabilidade condicional, é dado pela Equação 2.1, em que θ é uma lista contendo os parâmetros de interesse, $p(\theta/dados)$ é a distribuição *posteriori* dos parâmetros, $p(\theta)$ é a distribuição *priori* dos parâmetros, $p(dados/\theta)$ é a função de verossimilhança dos dados e $p(dados)$ é a probabilidade marginal dos dados, considerada uma constante normalizadora para que a *posteriori* mantenha as características de uma distribuição de probabilidade (integra em 1) (Van de Schoot and Miocević, 2020).

Ao contrário da abordagem clássica, que trata dos parâmetros de interesse como escalares ou vetores fixos passíveis de serem estimados apenas por via de seu valor mais provável (Murteira, 1995), a estatística bayesiana considera que os parâmetros de interesse comportam-se como variáveis aleatórias, e por isso suas estimativas são dadas na forma de funções de densidade de probabilidade, as quais contém seus próprios hiperparâmetros (Wagenmakers *et al.*, 2008). Dessa forma, à posteriori não incumbe somente determinar o valor mais provável para θ, como também evidenciar toda a incerteza a seu respeito (Murteira, 1995).

A *priori* consiste no conhecimento prévio existente a respeito dos parâmetros de interesse, cuja distribuição tem seus hiperparâmetros escolhidos pelo próprio pesquisador (Bernardo *et al.*, 2011). Essa distribuição pode (e preferencialmente deve) ser informativa, quando utiliza-se algum conhecimento pré-existente adquirido por meio de descobertas de estudos anteriores ou da opinião de especialistas para escolher o valor mais plausível, ou não-informativa, quando não há conhecimento prévio algum e é necessário assumir que todos os eventos do espaço paramétrico têm a mesma probabilidade de ocorrer (Lemoine, 2019).

Além disso, o pesquisador consegue refletir através de hiperparâmetros (como a variância, no caso de a distribuição escolhida ser a normal) o grau de incerteza com que o valor mais provável foi escolhido. Quanto menor o grau de incerteza, mais informativa será essa *priori*, exercendo uma maior influência sobre a construção da *posteriori*. A decisão acerca da distribuição que exercerá a função de *priori* exige muita cautela e experiência do pesquisador, afinal, a escolha de uma *priori* informativa inapropriada pode levar a resultados piores do que seriam com uma *priori* não informativa (Smid *et al.*, 2019).

A função de verossimilhança dos dados representa toda a informação que a amostra carrega a respeito dos parâmetros (Glickman and Van Dyk, 2007). Numericamente, é a probabilidade condicional da amostra dado o parâmetro θ, sendo representada pelo produtório (atendendo ao princípio da independência entre duas ou mais variáveis) das funções de probabilidade de cada elemento da amostra condicional a θ (Lindley, 1965).

A *posteriori*, na abordagem bayesiana, equivale à estimativa do parâmetro, na abordagem frequentista. Em outras palavras, consiste no resultado da atualização do conhecimento pré-existente com a nova informação a que se tem acesso, reportada por uma distribuição de probabilidade que caracteriza o comportamento dos parâmetros de interesse (Bernardo and Smith, 2009). Da mesma forma que ocorre com a *priori*, a interpretação da *posteriori* inclui o valor mais provável para o parâmetro, bem como o grau de incerteza vinculado a tal suposição.

Uma vez que a distribuição *posteriori* é resultado de uma combinação entre *priori* e função de verossimilhança, a sua forma e natureza será influenciada por ambas, ponderadas de acordo com vários fatores. Os dados amostrais, componentes primordiais da função de verossimilhança, determinam a influência que esta terá sobre o resultado final. A qualidade da *priori*, quanto ao nível de informação, também é um fator determinante para definir sua participação na obtenção da *posteriori* (Lenk and Orme, 2009).

Sinteticamente, o peso que a *priori* exercerá sobre a distribuição *posteriori* será diretamente proporcional ao quão informativa a *priori* é, e inversamente proporcional ao número de elementos que a amostra possui. Em outros termos, se a amostra for suficientemente grande e a *priori* for não-informativa, a função de verossimilhança terá uma influência maior na *posteriori*; se a amostra for pequena, mas a *priori* for informativa, a *posteriori* sofrerá mais influência da *priori*. Amostras grandes combinadas com

prioris informativas são o cenário ideal, enquanto que amostras pequenas combinadas com *prioris* não-informativas dificilmente proverão bons resultados (Lenk and Orme, 2009).

Desse modo, para amostras suficientemente grandes, é possível utilizar *prioris* não-informativas e ainda assim obter estimativas tão boas quanto seriam se utilizássemos o método frequentista. Em contrapartida, amostras pequenas fazem imprescindível que se considere uma distribuição informativa exercendo o papel de *priori* (Van de Schoot and Miocević, 2020).

Por essa razão, a abordagem bayesiana é amplamente explorada quando se trata de escassez de dados, ou ainda quando o número de variáveis preditoras é desproporcionalmente grande em relação ao número de observações – situação muito recorrente na seleção genômica, em que os marcadores moleculares (preditores) são expressivamente mais numerosos que os fenótipos (observações) (Wang *et al.*, 2018). Assim, é viável utilizar a inferência bayesiana na análise de pequenas amostras e obter estimativas bastante razoáveis para os parâmetros, desde que se tenha uma *priori* informativa e condizente com a realidade.

2.6 Métodos de Reamostragem

Boa parte das metodologias utilizadas na inferência estatística prometem estimar parâmetros com eficiência e credibilidade, desde que os dados atendam minimamente a uma série de pressuposições. Na prática, contudo, essa condição é frequentemente violada, seja devido a limitações na amostragem ou ao comportamento atípico dos dados. Ainda assim, em muitos casos é possível obter estimadores para os parâmetros de interesse, a despeito de uma qualidade questionável.

Um recurso válido para conferir maior confiabilidade a essas estimativas são os métodos de reamostragem (Chernick, 2012). Seu princípio é que, a partir de uma única amostra de tamanho qualquer, é possível extrair subconjuntos (chamados de subamostras) e estimar o mesmo parâmetro repetidas vezes utilizando cada um deles, constituindo as chamadas réplicas. Essa técnica permite avaliar a precisão de estimadores, discriminar *outliers*, constatar a ocorrência de *overfitting*, além de viabilizar a inferência estatística em amostras que não atendem as premissas básicas para a realização de determinados procedimentos (Yu, 2002).

2.6.1 Teste de Aleatorização

Também conhecido como Teste de Permutação, foi desenvolvido por Fisher (1935/1960), sendo largamente difundido anos mais tarde. O método consiste no rearranjo dos elementos da amostra original de todas as formas possíveis considerando que a hipótese nula é verdadeira. Por exemplo, se a H_0 sugere que não há diferença significativa entre dois tratamentos I e II, em teoria é possível "inverter"quantos e quaisquer valores observados de I e II sem alterar as características da amostra, partindo do princípio que elementos estatisticamente iguais são permutáveis (Yu, 2002).

Isso implica em um número de combinações que cresce exponencialmente a cada nova observação contida na amostra, fator que inviabilizava a realização do teste para grandes volumes de dados na época de Fisher, e que passou a prosperar posteriormente acompanhando os avanços computacionais (Good, 2013). Em amostras pequenas e fazendo uso de métodos de automatização, é praticável considerar todos os rearranjos existentes, já em amostras grandes pode ser necessário realizar simulações a fim de englobar o maior número de permutações possível. Ademais, o método oferece grande versatilidade quando se trata de delineamentos desbalanceados e dados faltantes (Hooton, 1991).

2.6.2 Validação cruzada

A validação cruzada simples foi inicialmente proposta por KURTZ (1948), sendo mais tarde aprimorada como validação cruzada dupla por MOSIER (1951), e posteriormente em sua versão múltipla por KRUS and FULLER (1982). O método mais utilizado atualmente é chamado de validação cruzada *k-fold*, cujo principal objetivo é verificar a replicabilidade de resultados, além de diagnosticar a ocorrência de sobreajuste.

Em problemas de classificação ou predição, é comum dispor de apenas uma amostra limitada que deve fazer-se suficiente para treinar um modelo e também testá-lo. Se os mesmos dados forem utilizados em ambas as etapas, ocorrerá *overfitting*, a acurácia do modelo será demasiadamente alta para a amostra com a qual ele foi treinado, porém baixa para novos dados de mesma natureza (YU, 2002). Ou seja, o modelo torna-se muito específico para determinado conjunto de dados, mas desprovido de qualquer utilidade prática (DIETTERICH, 1995).

O que se busca em estimadores de parâmetros é que não sejam tão específicos a ponto de apresentar viés e nem tão genéricos a ponto de não serem capazes de realizar uma predição com qualidade. Nenhum parâmetro é hábil para capturar a variação do acaso, logo a maximização da acurácia nem sempre reflete boas propriedades de $\hat{\theta}$.

A validação cruzada *k-fold* consiste em dividir aleatoriamente a amostra em k subgrupos disjuntos e de mesmo tamanho. A cada iteração, $k-1$ subconjuntos são utilizados para treinar o modelo (*training set*) e o restante para testá-lo (*validation set*). Após a realização de k iterações, em que todos os k subgrupos assumiram a posição de validação uma única vez, a acurácia do modelo é medida através da acurácia média das k iterações. Isso garante que os dados utilizados para treinar o modelo serão diferentes daqueles utilizados para testá-lo, conferindo uma maior confiabilidade na sua acurácia na medida em que evita o sobreajuste (BERRAR, 2019).

2.6.3 *Jackknife*

O método *jackknife*, assim nomeado anos depois de sua criação, foi introduzido por QUENOUILLE (1949) e posteriormente explorado por TUKEY (1958). No primeiro momento, Tukey dedicou-se ao método com a finalidade de investigar a influência de *outliers* na estimação de parâmetros a partir de subconjuntos da amostra (YU, 2002).

Também referida como *leave-one-out test*, a técnica consiste em remover um único elemento por vez da amostra original, e utilizar os dados restantes para compor uma amostra *jackknife*, a partir da qual é estimado o parâmetro de interesse correspondente, chamado de réplica *jackknife*. O objetivo é repetir esse procedimento para cada elemento da amostra, e por fim estimar o parâmetro através da média de todas as réplicas. Dessa forma, ao utilizar uma amostra com k elementos, é possível formar k amostras *jackknife* com $k-1$ elementos cada, além das correspondentes k réplicas *jackknife* (SHAO and TU, 2012).

Trata-se de um algoritmo de fácil implementação e possui um número fixo de iterações, igual ao número de elementos da amostra original. Além do mais, o método permite ainda calcular a estimativa *jackknife* do erro padrão e a estimativa *jackknife* do viés, propriedades importantes para se avaliar estimadores de modo geral (YU, 2002).

2.6.4 *Bootstrap*

O método *bootstrap* foi criado por Efron (1992) e posteriormente desenvolvido por Efron and Tibshirani (1994), já com o intuito de aplicar a técnica em problemas de inferência. Assim como o *jackknife*, é utilizado com o objetivo de medir o erro padrão e o viés de estimadores (Johnson, 2001), com a diferença de que o *bootstrap* permite um número maior de réplicas e, portanto, de iterações (Yu, 2002).

Partindo-se do mesmo princípio do *jackknife*, são geradas várias amostras *bootstrap* diferentes a partir de uma mesma amostra original. Essas amostras, por sua vez, são utilizadas para calcular estimativas (réplicas) para o parâmetro de interesse. A distinção dos métodos se faz presente na etapa de obtenção das amostras, que no *bootstrap* mantêm o mesmo tamanho da amostra original. O conteúdo de cada amostra *bootstrap* é proveniente de um sorteio aleatório com reposição, a partir dos dados amostrais (Dixon, 2006). Logo, ao utilizar umas amostra com k elementos, é possível formar uma quantidade ilimitada de amostras *bootstrap*, cada uma com k elementos, a partir dos quais é possível obter uma réplica *bootstrap* (Efron and Tibshirani, 1994).

3 MATERIAL E MÉTODOS

3.1 Estudo de simulação

3.1.1 A população inicial

Para este trabalho, foi realizado um estudo com dados simulados a fim de avaliar e comparar o desempenho dos métodos frequentista e bayesiano na estimação de um parâmetro qualquer. Todos os procedimentos, desde as simulações até as análises, foram executados utilizando o software livre R (R Core Team, 2017).

Inicialmente, foram simulados N valores representando uma população inteira com uma determinada característica de interesse quantitativa contínua descrita através de uma variável aleatória seguindo uma distribuição normal, uma vez que a maior parte das características quantitativas do âmbito agronômico apresentam distribuição aproximadamente normal. Também foram escolhidos, de forma totalmente arbitrária, valores de referência para parâmetros média (μ) e variância (σ^2) dessa população.

Em seguida, a partir dessa população foi extraída, através de um sorteio, uma amostra de tamanho n, sobre a qual foram realizadas a inferência clássica e bayesiana, dois métodos já conhecidos na literatura para a estimação dos parâmetros da população global. O objetivo desse estudo de simulação foi avaliar qual método apresentou um melhor performance, comparando as estimativas obtidas com o real valor atribuído ao parâmetro.

A população inicial simulada continha $N = 10^6$ indivíduos, quantidade grande o bastante para que os dados gerados assumissem um comportamento bem característico de uma distribuição normal. Os valores que foram assumidos pelos parâmetros média e variância dessa população para um variável de interesse qualquer foram, respectivamente, $\mu = 10$ e $\sigma^2 = 4$, ambos escolhidos de forma completamente aleatória.

Além disso, foram testados 4 diferentes tamanhos amostrais, com n variando entre os valores 5, 10, 20 e 30 elementos. A escolha dos valores de n se deu por abranger valores considerados pequenos em relação ao que é comum observar a respeito de tamanho amostral na prática. Finalmente, com o objetivo de avaliar o efeito de diferentes *prioris* no resultado final obtido no caso da inferência bayesiana, foram consideradas 3 diferentes *prioris*: menos informativa (baseada em 2 experimentos prévios), intermediária (baseada em 6 experimentos prévios) e mais informativa (baseada em 10 experimentos prévios). Cada experimento prévio gerou uma amostra com o mesmo número de elementos da atual, simulando que o método de amostragem foi mantido em todas as situações.

3.1.2 Inferência clássica

Na inferência clássica, a estimação pontual de parâmetros se dá por estimadores, ou seja, estatísticas ou funções dos dados que estão sendo estudados. Uma vez que a amostra seja representativa da população, suficientemente grande e que o estimador apresente as propriedades de suficiência, consistência e eficiência e ausência de viés, é esperado que a estimativa se aproxime muito do real valor do parâmetro populacional.

O melhor estimador não viesado para a média populacional é a própria média amostral, representada pela Equação 3.1, em que $\hat{\mu}$ é a estimativa de μ, x_i é o i-ésimo elemento da amostra, e n é o tamanho da amostra.

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.1}$$

Quanto à variância, o melhor estimador não viesado é dado pela Equiação 3.2, em que $\hat{\sigma^2}$ é a estimativa de σ^2, x_i é o i-ésimo elemento da amostra, n é o tamanho da amostra e $\overline{x}$ é a média aritmética de seus elementos.

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \tag{3.2}$$

Para cada parâmetro, foram obtidas oito estimativas, correspondentes aos quatro diferentes tamanhos amostrais testados multiplicado pelos dois métodos de reamostragem utilizados. Além de proporcionar as estimativas para os parâmetros de interesse, tanto o método *jackknife* como o *bootstrap* permitem o cálculo das estimativas do erro padrão e do viés para média e variância.

A Equação 3.3 apresenta o cálculo da estimativa *jackknife* do erro padrão, onde n é o tamanho amostral, $\theta_{(i)}$ é a i-ésima réplica *jackknife*, e $\overline{\hat{\theta}}$ é a estimativa *jackknife* de θ.

$$\hat{ep}_j = \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{(i)} - \overline{\hat{\theta}})^2} \tag{3.3}$$

Paralelamente, a Equação 3.4 mostra o cálculo da estimativa *bootstrap* do erro padrão, em que B é número de amostras *bootstrap*, $\theta_{(j)}$ é a j-ésima réplica *jackknife*, e $\overline{\hat{\theta}}$ é a estimativa *jackknife* de θ.

$$\hat{ep}_b = \sqrt{\frac{1}{B-1} \sum_{j=1}^{B} (\hat{\theta}_{(j)} - \overline{\hat{\theta}})^2} \tag{3.4}$$

Finalmente, a Equação 3.4 traz o cálculo da estimativa do viés, onde n é o tamanho amostral, $\overline{\hat{\theta}}$ é a estimativa *jackknife* ou *bootstrap* do parâmetro, e $\hat{\theta}$ é a estimativa do parâmetro utilizando todos os elementos da amostra.

$$\hat{vies} = (n-1)(\overline{\hat{\theta}} - \hat{\theta}) \tag{3.5}$$

Vale salientar que, enquanto as estimativas do erro padrão para determinado parâmetro são sempre positivas devido à raiz quadrada da expressão, a estimativa do viés pode sim apresentar valores negativos.

3.1.3 Inferência bayesiana

Na inferência bayesiana, a estimação de um parâmetro resulta em uma distribuição indicando seus valores mais prováveis, chamada *posteriori*. Para aplicar o Teorema de Bayes e obter a *posteriori*, é necessário construir a função de verossimilhança dos dados a partir da amostra e determinar uma distribuição *priori*.

Assim como na inferência clássica, o número de elementos da amostra oscilou entre os valores 5, 10, 20 e 30. A respeito da *priori*, como tratava-se de uma simulação com valores arbitrários para média e variância populacionais e não havendo valores de referência na literatura, a *priori* foi construída da seguinte forma: foram realizadas 2, 6 ou 10 amostragens a partir da população total, representando assim a realização de experimentos realizados no passado. Em cada cenário, o tamanho das amostras prévias foi mantido para a amostra atual.

Para aplicar o método, foi utilizado o pacote *stan* (Stan Development Team, 2022), específico para análises bayesianas, disponível para a linguegem R. Em sua sintaxe, é necessário construir um modelo definindo as propriedades dos parâmetros a serem estimados, tais como condições de existência e distribuições de probabilidade dos hiperparâmetros, e suas respectivas *prioris*.

Sucintamente, para cada amostragem prévia representando um experimento anterior foram calculadas a média e variância amostrais. Em seguida, notou-se que, quando avaliados conjuntamente todos os valores que cada parâmetro assumiu ao longo dos experimentos anteriores, foi observado um comportamento semelhante a uma distribuição normal. Dessa forma, a *priori* da média seguiu distribuição normal com média e variância calculadas a partir das estimativas de média das amostragens prévias, e a *priori* da variância seguiu distribuição normal com média e variância calculadas a partir das estimativas de variância dos experimentos anteriores.

Esses valores foram então informados no modelo *stan*, e a análise executada, resultando em um grande número de valores representando a distribuição *posteriori* dos parâmetros média e variância. Assim como para a abordagem frequentista, foram utilizados os métodos de reamostragem *jackknife* e *bootstrap*, servindo novamente como indicadores de qualidade dos processos de estimação.

3.2 Aplicação prática

3.2.1 Os dados reais

Este trabalho utilizou dados reais coletados pelo Departamento de Ciências Biológicas da Escola Superior de Agricultura Luiz de Queiroz (ESALQ). As amostras referem-se a duas espécies típicas do Cerrado brasileiro ocorrendo em três diferentes áreas: Cerrado natural, Cerrado em regeneração e sub-bosque de *Pinus*. A coleta do material foi realizada na Estação Ecológica de Santa Bárbara (EEcSB), na cidade de Águas de Santa Bárbara, estado de São Paulo.

As espécies utilizadas foram *Duguetia furfuracea* (A.St.-Hil.) Saff., uma planta arbustiva da família Annonaceae popularmente conhecida como araticum, marolinho-do-cerrado ou pinha-de-guará; e *Licania humilis* Cham. & Schltdl., árvore da família Chrysobalanaceae, mais conhecida como fruta-de-ema. Para as áreas de Cerrado natural e Cerrado em regeneração, foram amostrados ramos de cinco indivíduos de cada espécie, e para o sub-bosque de *Pinus*, apenas 3 indivíduos.

As variáveis de interesse são atributos anatômicos foliares associados à adaptação das espécies, e foram mensuradas através de técnicas histológicas convencionais. Para este trabalho, serão considerados os seguintes atributos: espessura da cutícula da face superior, espessura da parede celular periclinal externa, altura da célula epidérmica superior, espessura do mesofilo e altura da célula epidérmica inferior. A partir de uma mesma folha foram medidas todas as variáveis citadas, preservando a informação de possíveis correlações entre atributos de um mesmo indivíduo.

Além disso, constatou-se que diferentes folhas de um mesmo indivíduo não apresentaram correlação significativa entre si para os atributos considerados, justificando a utilização de cada folha como sendo uma observação diferente. Isso decorre do fato de que as características estudadas estão muito mais subordinadas à posição da planta em que se encontram do que ao material genético individual propriamente dito.

3.2.2 Análise

Embora tenha sido evidenciada uma melhor aptidão da abordagem bayesiana para a estimação de um parâmetro a partir de uma amostra com poucos elementos utilizando dados simulados, ambos os métodos frequentista e bayesiano foram novamente utilizados para estimar as médias populacionais para cada uma das características consideradas, a fim de corroborar a ideia de que as diferentes abordagens levam a resultados significativamente distintos.

Foram consideradas amostras de ramos de *L. humilis*, sendo 5 indivíduos na área de Cerrado natural, 5 indivíduos na área de Cerrado em regeneração e 3 indivíduos no sub-bosque de *Pinus*. Para a inferência clássica, o estimador utilizado foi a média amostral descrita anteriormente, sendo estimada uma média para cada um dos três ambientes.

Para a inferência bayesiana, as mesmas amostras de *L. humilis* foram consideradas. Com relação à *priori*, não há na literatura relatos dos mesmos atributos foliares dessa espécie especificamente nos três tipos de vegetação considerados, o que seria considerada uma *priori* bastante informativa. Alternativamente, optou-se por utilizar os dados de *D. furfuracea* como *priori*, pois apesar de serem de espécies diferentes, os dados foram coletados nas mesmas áreas, mesma época e com atributos medidos segundo os mesmos critérios dos dados de *L. humilis*, sendo assim a melhor informação que se tem com relação ao fenômeno estudado.

Assim como no estudo de simulação, todas as análises foram realizadas utilizando o software R (R Core Team, 2017). Para a estimação dos parâmetros via abordagem bayesiana, foi utilizado o pacote *stan* (Stan Development Team, 2022). A *priori* da média foi definida conforme o distribuição dos dados de *D. furfuracea*, seguindo distribuição normal com média e variância calculadas a partir desses dados. Dessa forma, foi realizada uma estimação para cada combinação de variável resposta e local de ocorrência (por exemplo, a média de espessura da cutícula superior para indivíduos localizados no Cerrado natural).

Em seguida, para cada um dos cinco atributos foliares de interesse, foi realizada uma Análise da Variância (ANOVA) com nível de significância para o teste F definido em 5%, a fim de avaliar as seguintes hipóteses:

$$\begin{cases} H_0 : \mu_C = \mu_R = \mu_P \\ H_a : \text{pelo menos duas médias diferem entre si.} \end{cases}$$

em que H_0 é a hipótese nula, H_a é a hipótese alternativa, μ_C, μ_R e μ_P são, respectivamente, as médias de indivíduos ocorrendo no Cerrado natural, no Cerrado em regeneração e no sub-bosque de *Pinus*, para determinada caraterística de interesse. Uma vez rejeitada a hipótese de nulidade para um atributo, foi realizado um teste de Tukey com nível de significância de 0,05 para comparar as médias duas a duas, utilizando a função *HSD.test* do pacote *agricolae* (de Mendiburu, 2021) disponível para a linguagem R.

3.3 Regressão

Uma possível alternativa para quando a escassez de dados se dá em decorrência de uma variável de difícil mensuração é utilizar outra variável que possa ser medida de forma mais simples e que seja altamente correlacionada com a de interesse, de forma a predizê-la. Para testar a técnica no conjunto de dados de *L. humilis*, a primeira etapa foi realizar o cálculo da correlação entre todas as variáveis a fim de

identificar quais poderiam ser utilizadas como preditora e predita. As duas variáveis mais correlacionadas entre si, em módulo, foram selecionadas, sendo uma delas escolhida aleatoriamente para ser a variável de interesse.

Em seguida, utilizando a totalidade dos dados, foram realizadas regressões polinomiais a fim de caracterizar o comportamento que uma variável exerce sobre a outra através da escolha do grau do polinômio. O princípio do método consiste em aumentar progressivamente o grau do polinômio até verificar que os desvios de regressão deixam de ser significativos na ANOVA. Quando isso ocorre, entende-se que o comportamento da característica escolhida como variável resposta pode ser representado através de uma função da outra característica.

A partir dessa informação, tem início o processo de predição da variável de interesse. Com o intuito de monitorar a ocorrência indesejada de sobreajuste, foi utilizada a validação cruzada 5^{th} *fold*. Para cada grupo de treinamento, foram obtidos os coeficientes da função através da regressão polinomial de grau tido como mais adequado na etapa anterior, e utilizados para predizer a variável de interesse. A ideia é que, se para todos os cinco grupos de treinamento, o polinômio considerado mais apropriado for de mesmo grau, há um indício de que não há sobreajuste, pois um único modelo conseguiu capturar o comportamento geral dos dados.

Por fim, realizou-se uma análise da variância a fim de avaliar a influência das três localidades (Cerrado Natural, Cerrado em Regeneração e Sub-bosque de *Pinus*) na variável de interesse predita, sendo este o objetivo inicial do trabalho. Tendo disponíveis os dados originais dessa variável, aqueles que foram utilizados para o treinamento do modelo, foi realizada outra análise da variância com a finalidade de comparar os resultados obtidos através dos valores reais e preditos. Além disso, uma vez que a ANOVA aponte diferença significativa entre as localidades, faz-se necessário realizar um teste de comparação de médias, sendo escolhido o Teste de Tukey utilizando a biblioteca *agricolae*.

4 RESULTADOS

4.1 Estudo de simulação

As tabelas 4.1 a 4.4 trazem os valores da estimativa do erro padrão e da estimativa do viés, tanto para a média como para a variância, utilizando os métodos de reamostragem *jackknife* e *bootstrap*. Nas linhas, a letra n representa os diferentes tamanhos amostrais. Os termos "pouco informativa", "intermediária"e "muito informativa"utilizados para descrever a *priori* fazem referência à informação prévia baseada em 2, 6 e 10 experimentos, respectivamente.

Tanto a estimativa do erro padrão como a estimativa do viés para determinado parâmetro têm como principal objetivo determinar a incerteza em torno da estimativa, de forma que um estimador de alta qualidade está associado a valores baixos, em módulo, para ambos os indicadores. Isso porque, dependendo do método de estimação utilizado, valores atípicos podem interferir demasiadamente no valor final da estimativa, e ao utilizar esses métodos de reamostragem, isso é refletido na medida em que os valores das réplicas mostram-se muito distintos.

Em geral, é esperado que conforme o número de elementos por amostra cresce, a estimativa do erro padrão seja reduzida. O que foi possível perceber com base nas tabelas é que, independentemente do método de reamostragem ser *jackknife* ou *bootstrap*, a diferença entre os indicadores para os diferentes tamanhos amostrais foi mais expressiva quando foi utilizada a *priori* mais informativa, no caso da análise bayesiana.

Como era esperado, as estimativas dos vieses da média e da variância obtidas através da análise frequentista utilizando o método de reamostragem *jackknife* foram todas iguais a zero, o que não ocorreu para as estimativas dos vieses da média e da variância utilizando *bootstrap*.

Isso se deve ao fato de que o método *jackknife* utiliza todos os elementos da amostra exatamente o mesmo número de vezes, de forma que a média aritmética das réplicas *jackknife* naturalmente coincide com a média aritmética dos elementos da amostra, o que também é utilizado como estimador do parâmetro média populacional na análise frequentista. Em contrapartida, o método *bootstrap* utiliza subamostras obtidas por meio de sorteios com reposição de elementos, portanto, não conservam necessariamente as mesmas proporções da amostra original, o que justifica essa diferença.

Tabela 4.1. Estimativa do erro padrão e estimativa do viés para média populacional ($\hat{\mu}$) utilizando o método de reamostragem *jackknife* para 4 diferentes tamanhos amostrais e 3 diferentes *prioris*

			Bayesiana			Frequentista
			Priori			
			Pouco Informativa	Intermediária	Informativa	
Estimativa do Erro Padrão	n	5	0,026	0,369	0,489	0,791
		10	0,025	0,164	0,117	0,589
		20	0,001	0,014	0,017	0,477
		30	0,044	0,009	0,025	0,360
Estimativa do Viés	n	5	-0,013	0,523	0,136	0
		10	0,029	0,055	0,036	0
		20	-0,014	-0,049	0,120	0
		30	0,040	-0,021	-0,055	0

Tabela 4.2. Estimativa do erro padrão e estimativa do viés para variância populacional ($\hat{\sigma^2}$) utilizando o método de reamostragem *jackknife* para 4 diferentes tamanhos amostrais e 3 diferentes *prioris*

			Bayesiana			Frequentista
			Priori			
			Pouco Informativa	Intermediária	Informativa	
Estimativa do Erro Padrão	n	5	1,707	0,395	1,074	2,434
		10	0,371	0,650	0,482	1,455
		20	0,252	0,287	0,232	1,174
		30	0,298	0,291	0,283	1,118
Estimativa do Viés	n	5	1,688	0,631	0,257	0
		10	2,989	-0,455	-0,013	0
		20	0,128	0,133	-0,575	0
		30	0,342	-0,905	0,154	0

Tabela 4.3. Estimativa do erro padrão e estimativa do viés para média populacional ($\hat{\mu}$) utilizando o método de reamostragem *bootstrap* para 4 diferentes tamanhos amostrais e 3 diferentes *prioris*

			Bayesiana			Frequentista
			Priori			
			Pouco Informativa	Intermediária	Informativa	
Estimativa do Erro Padrão	n	5	1,614	0,281	0,849	1,018
		10	0,324	0,043	0,103	0,781
		20	0,029	0,173	0,031	0,471
		30	0,003	0,088	0,038	0,372
Estimativa do Viés	n	5	8,039	0,021	1,655	-1,406
		10	-0,095	-0,143	-0,005	-1,509
		20	0,500	0,704	0,312	1,730
		30	-0,100	-0,547	-0,266	1,223

Tabela 4.4. Estimativa do erro padrão e estimativa do viés para variância populacional ($\hat{\sigma^2}$) utilizando o método de reamostragem *bootstrap* para 4 diferentes tamanhos amostrais e 3 diferentes *prioris*

			Bayesiana			Frequentista
			Priori			
			Pouco Informativa	Intermediária	Informativa	
Estimativa do Erro Padrão	n	5	0,038	1,255	0,426	2,815
		10	0,485	0,383	0,890	1,547
		20	0,221	0,426	0,222	1,272
		30	0,047	0,207	0,029	0,768
Estimativa do Viés	n	5	-0,112	-2,088	-3,419	6,036
		10	-0,411	0,015	-3,984	6,725
		20	-1,632	-0,721	0,068	-3,952
		30	-0,066	0,976	-4,713	-0,509

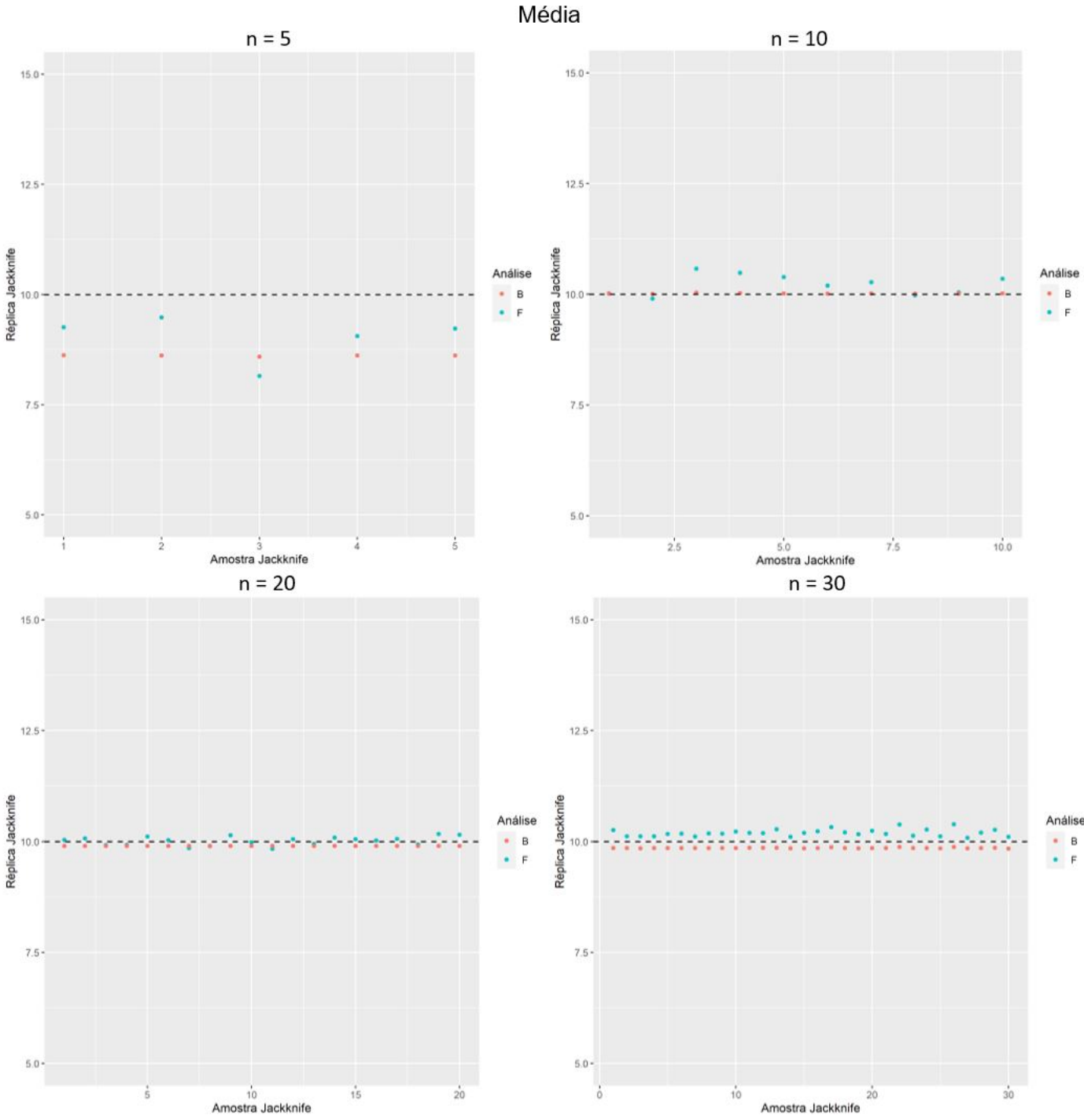

Figura 4.1. Estimativas *Jackknife* para média utilizando *priori* pouco informativa

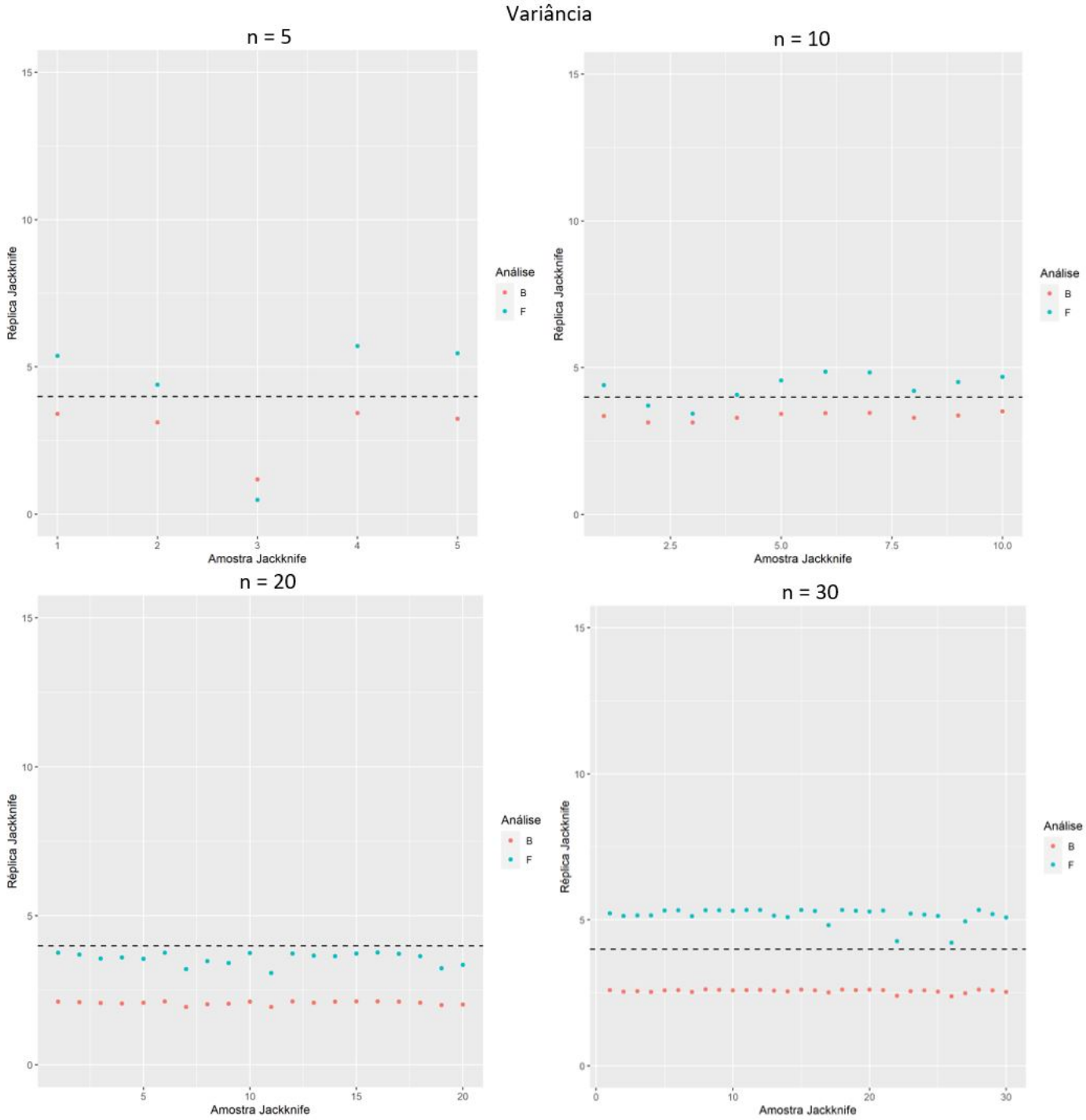

Figura 4.2. Estimativas *jackknife* para variância utilizando *priori* pouco informativa

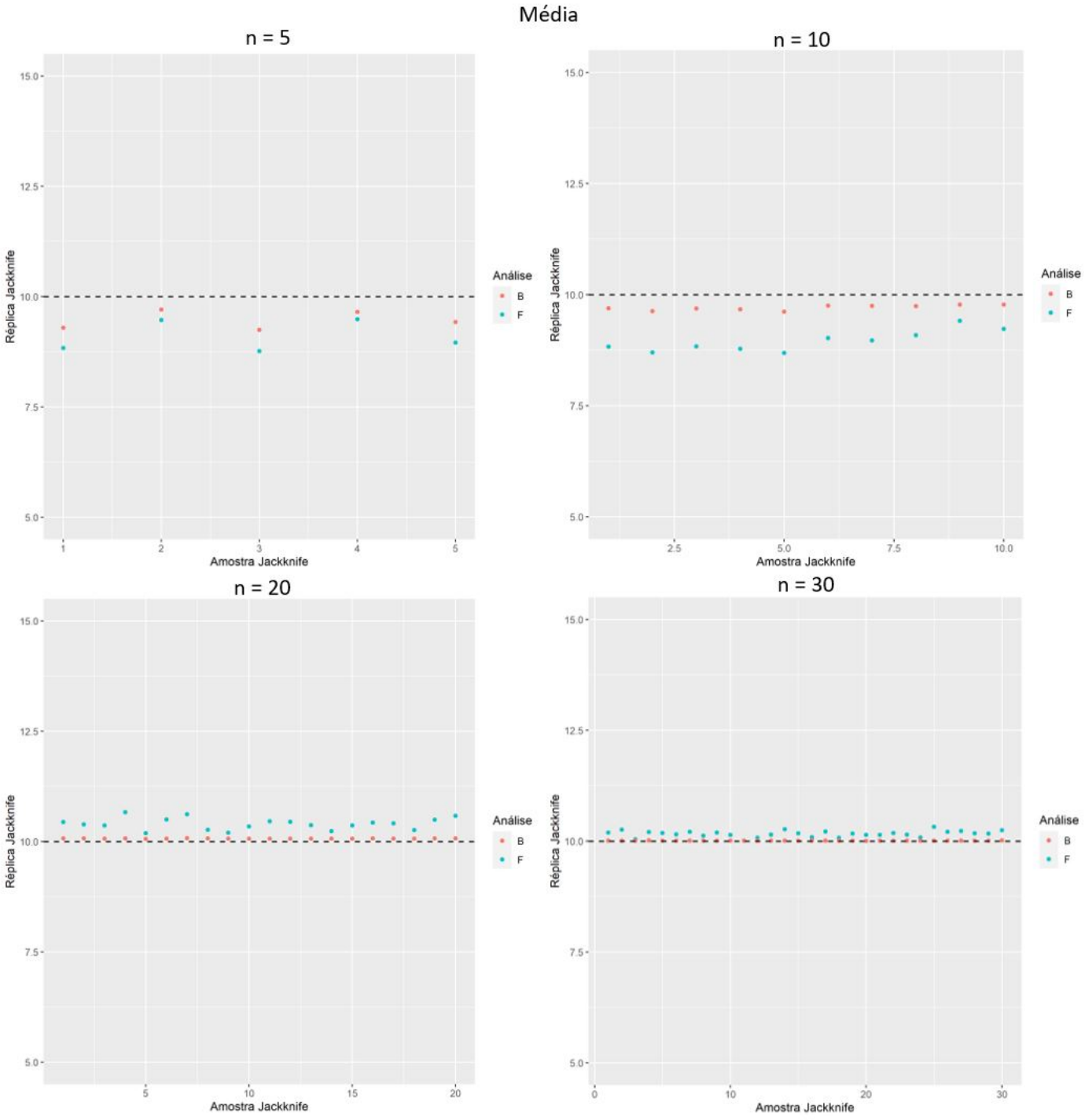

Figura 4.3. Estimativas *jackknife* para média utilizando *priori* intermediária

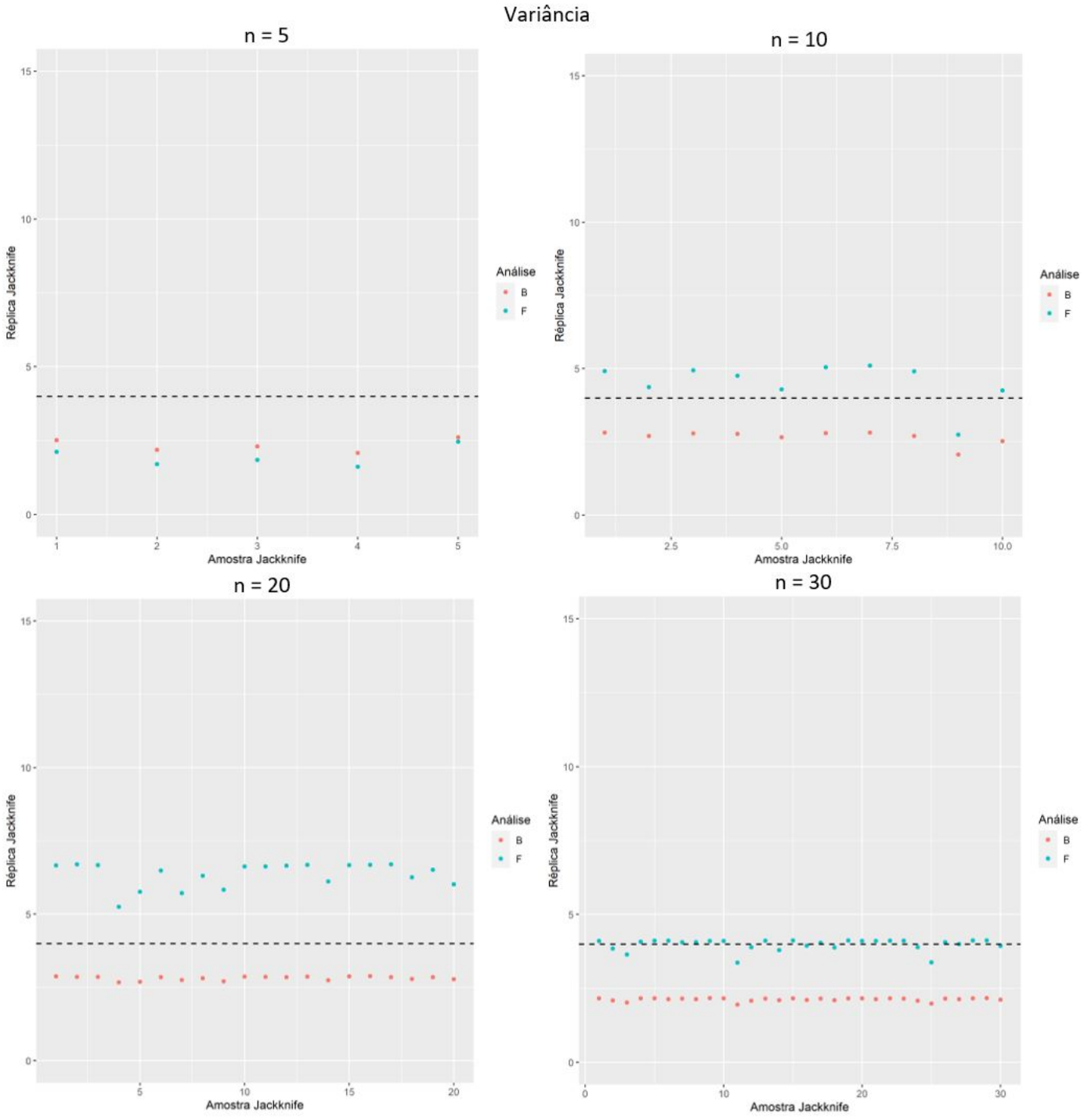

Figura 4.4. Estimativas *jackknife* para variância utilizando *priori* intermediária

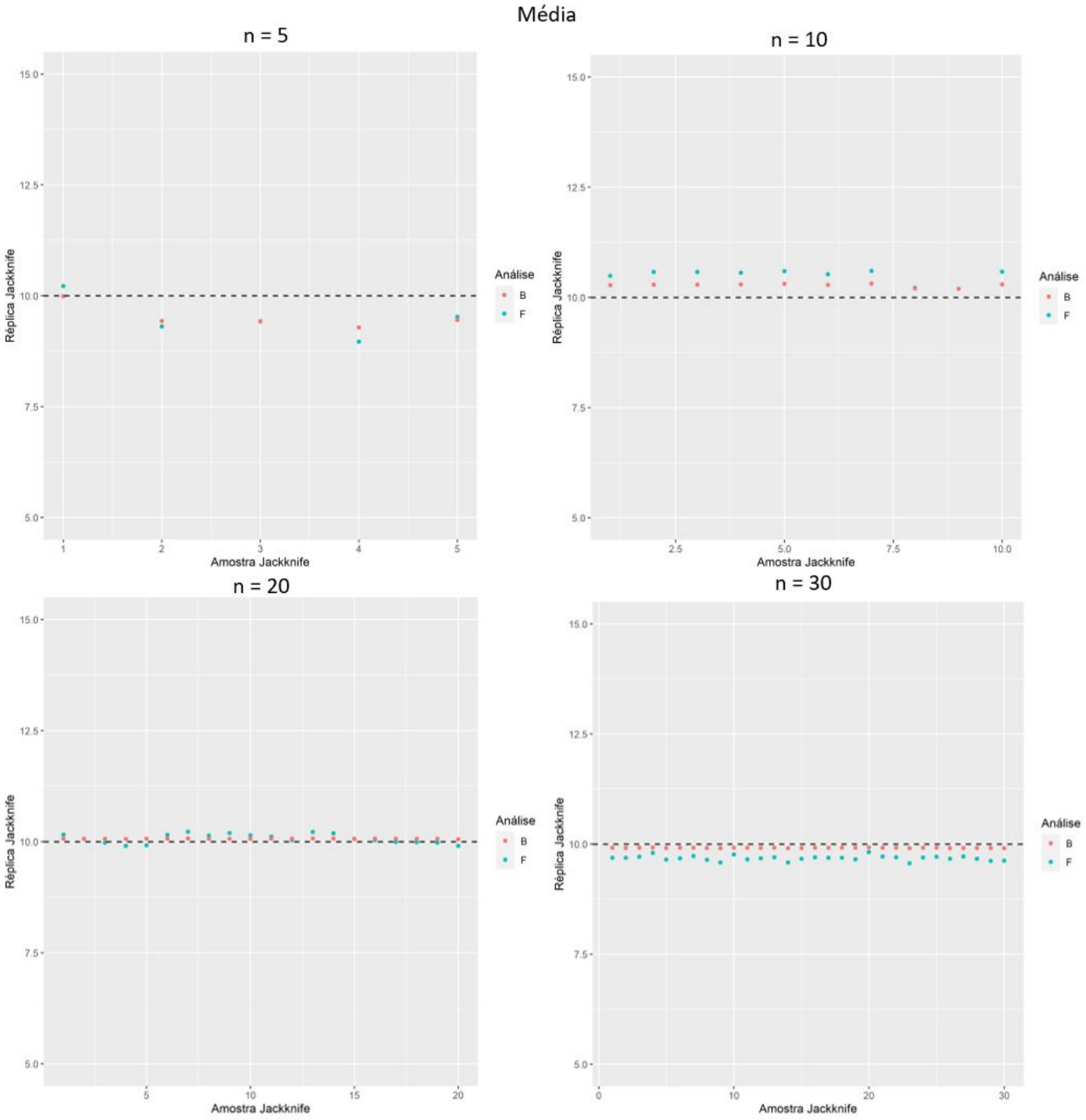

Figura 4.5. Estimativas *jackknife* para média utilizando *priori* informativa

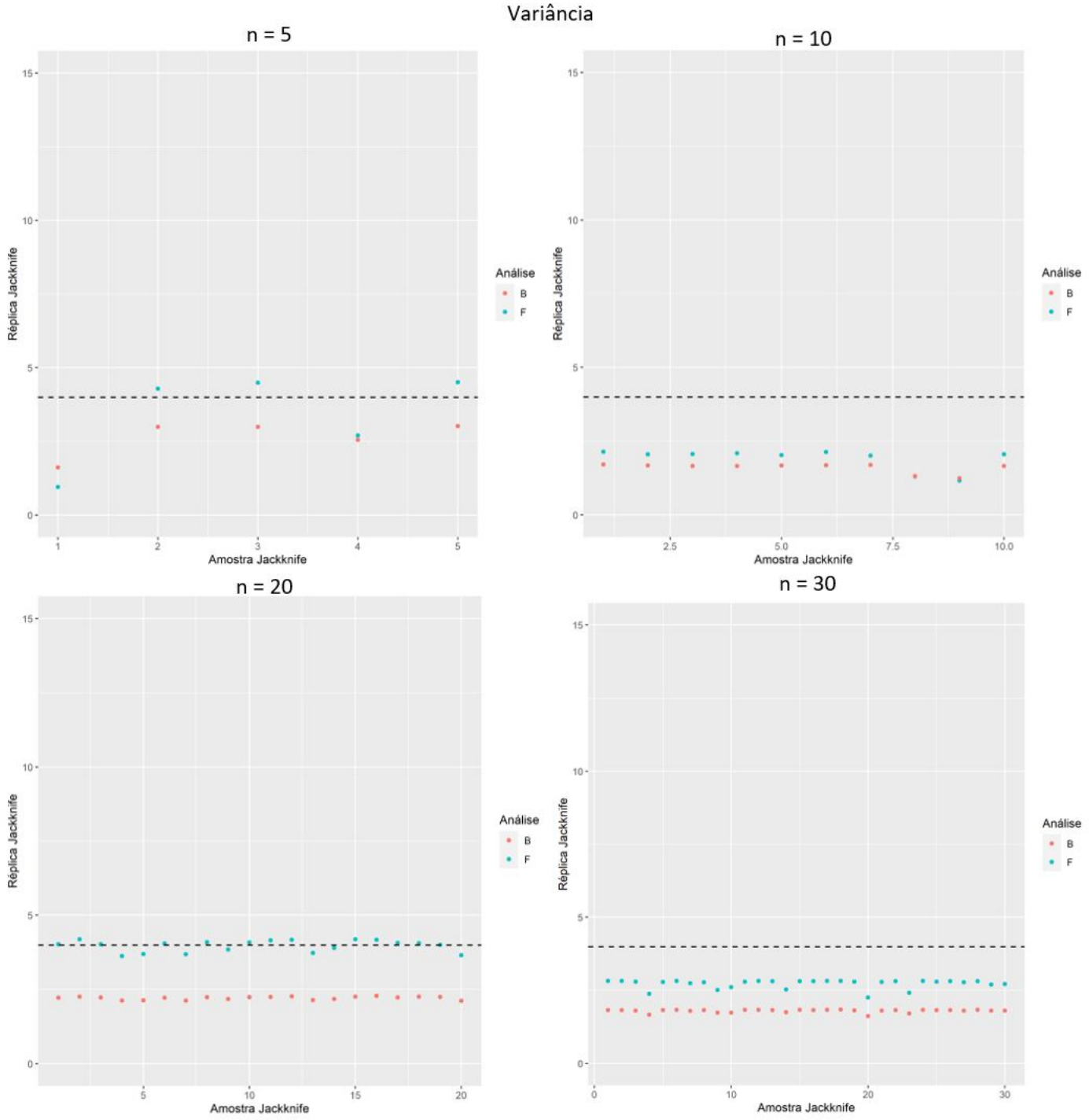

Figura 4.6. Estimativas *jackknife* para variância utilizando *priori* informativa

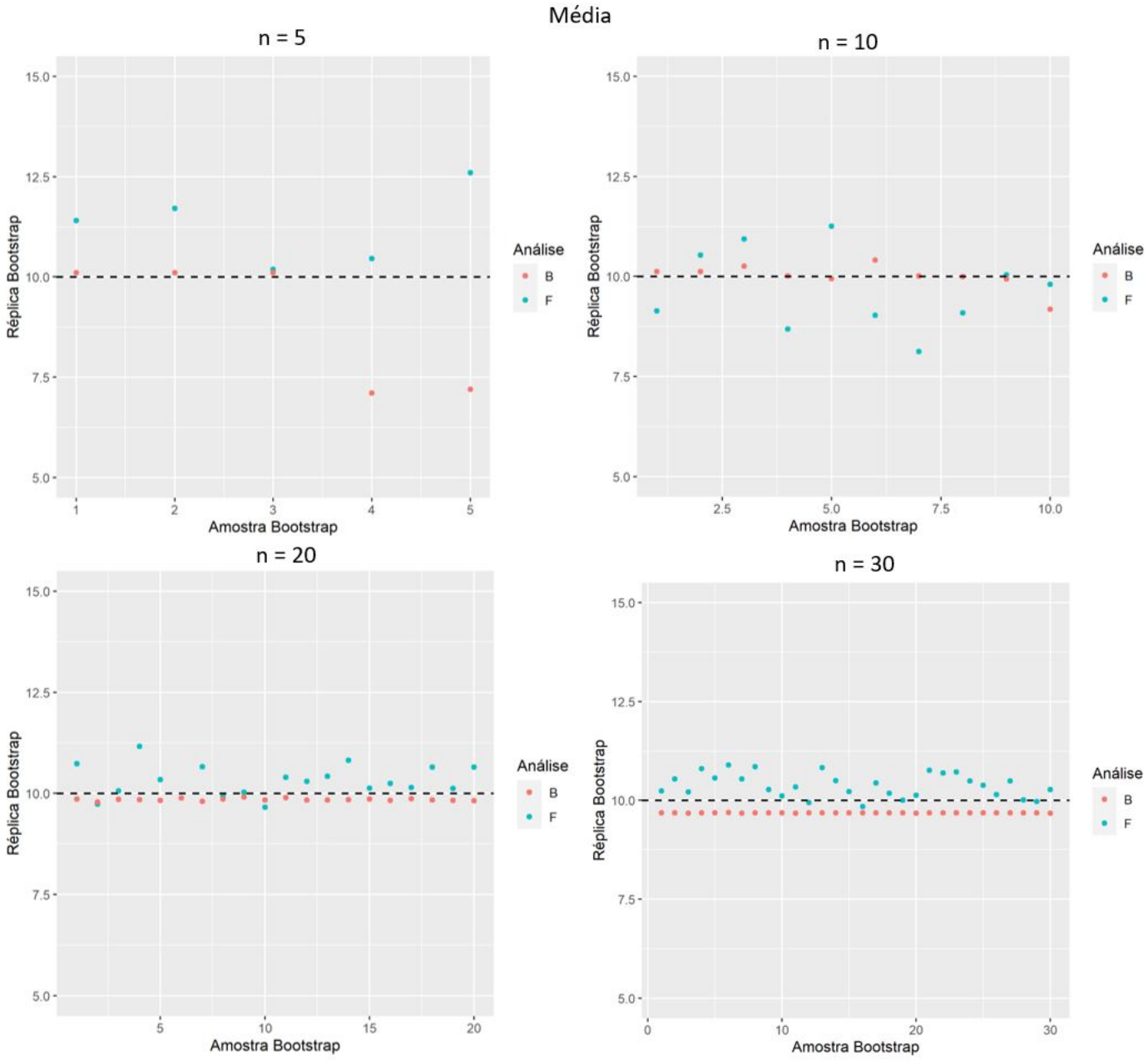

Figura 4.7. Estimativas *bootstrap* para média utilizando *priori* pouco informativa

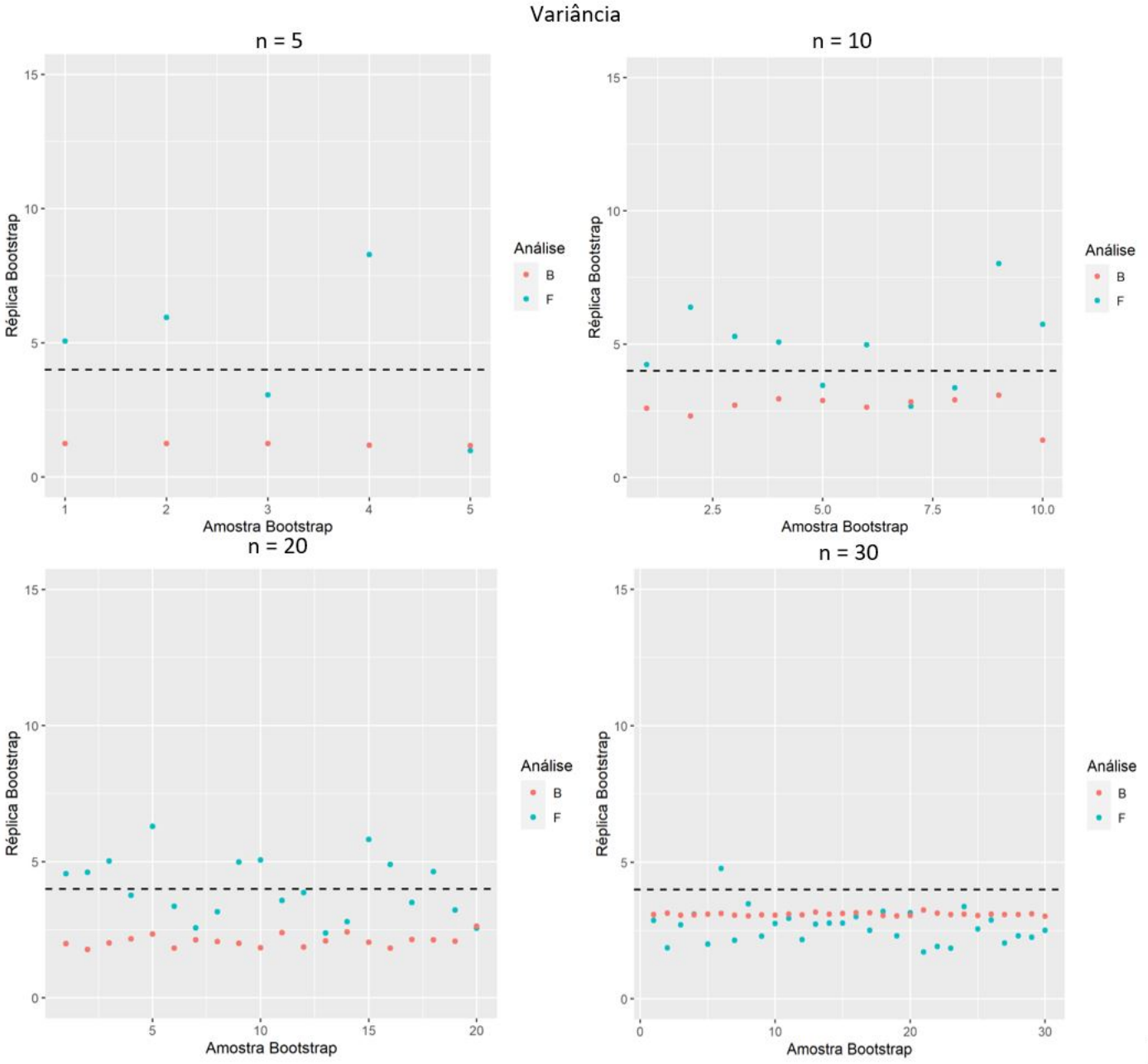

Figura 4.8. Estimativas *bootstrap* para variância utilizando *priori* pouco informativa

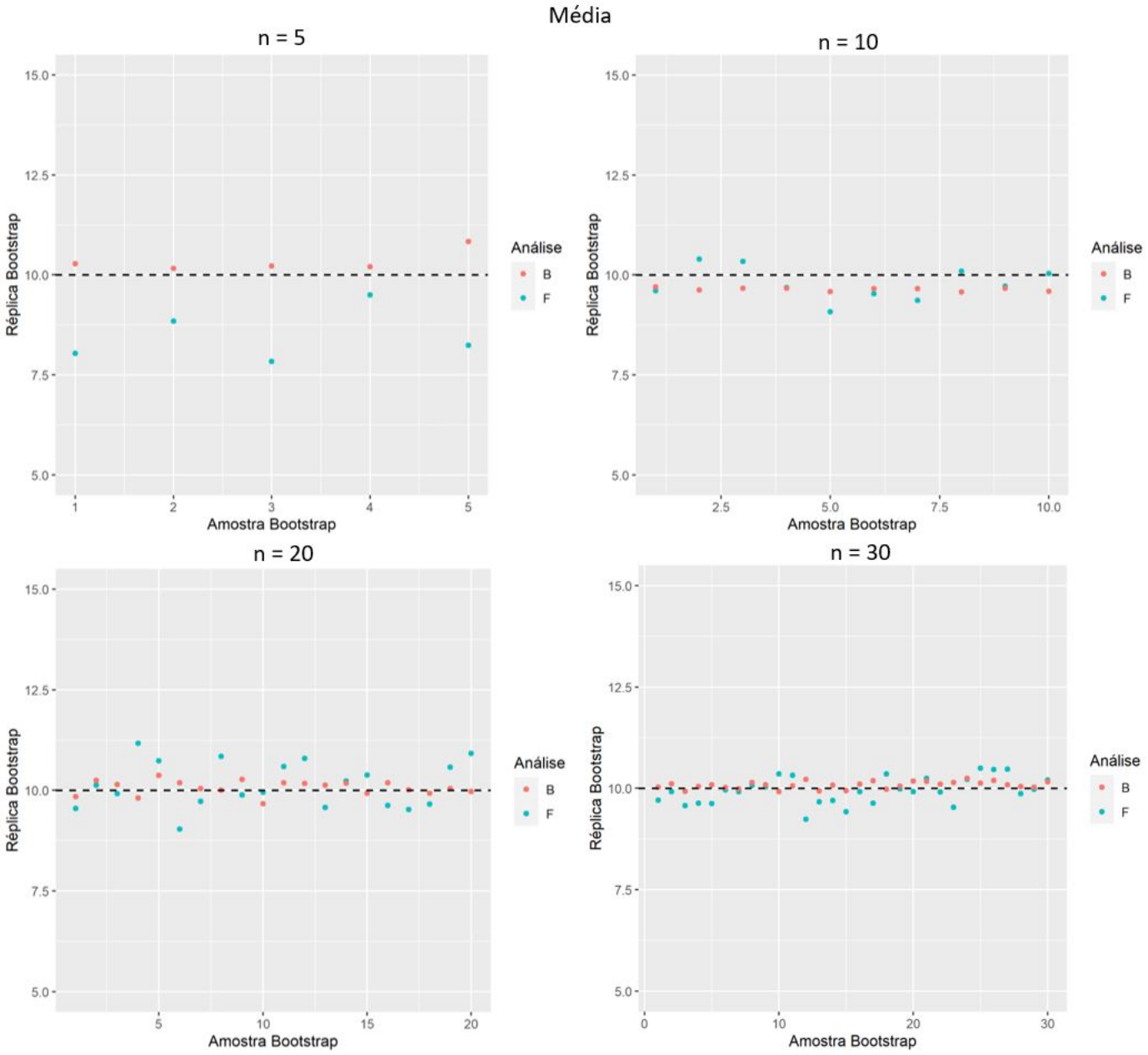

Figura 4.9. Estimativas *bootstrap* para média utilizando *priori* intermediária

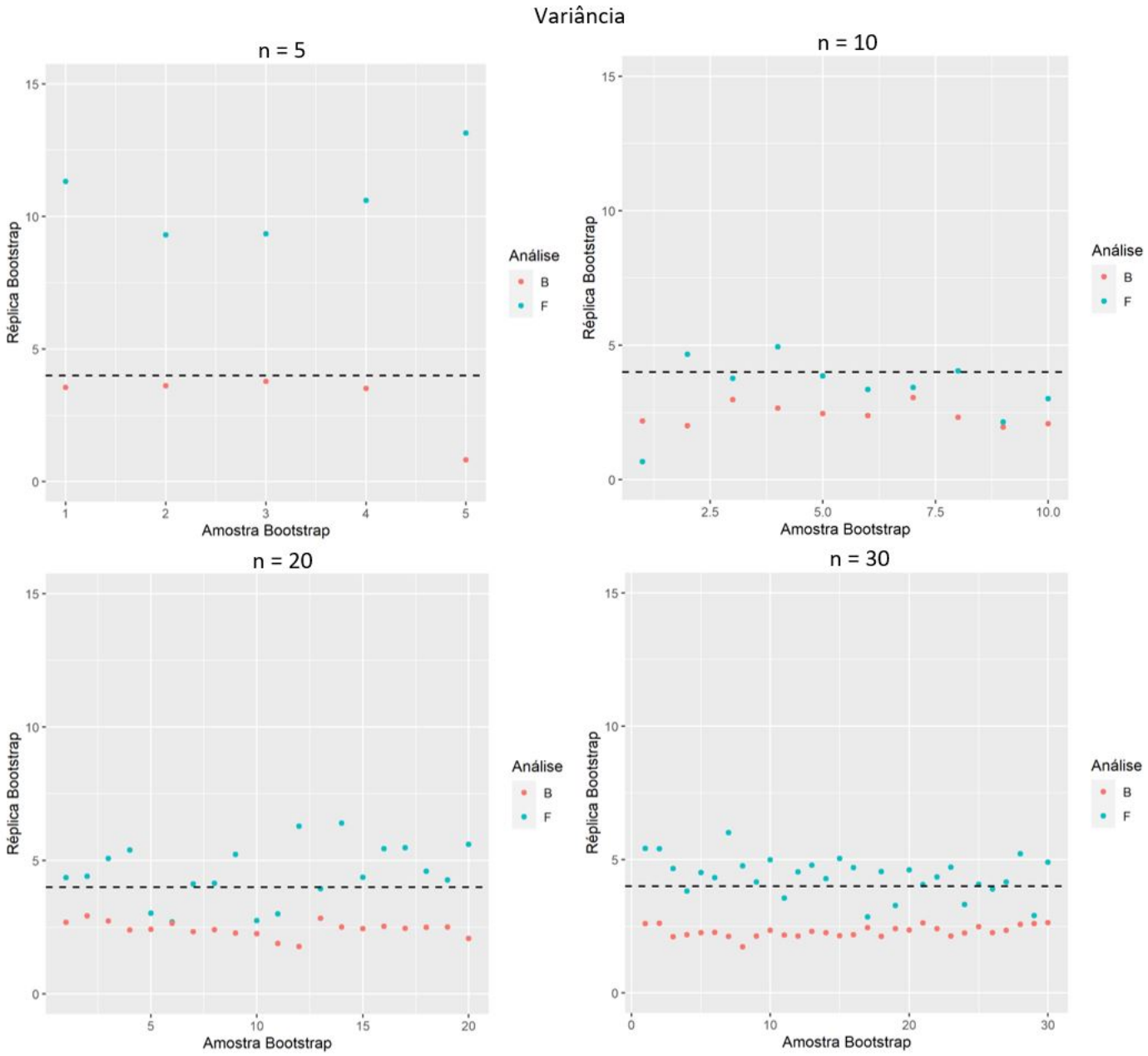

Figura 4.10. Estimativas *bootstrap* para variância utilizando *priori* intermediária

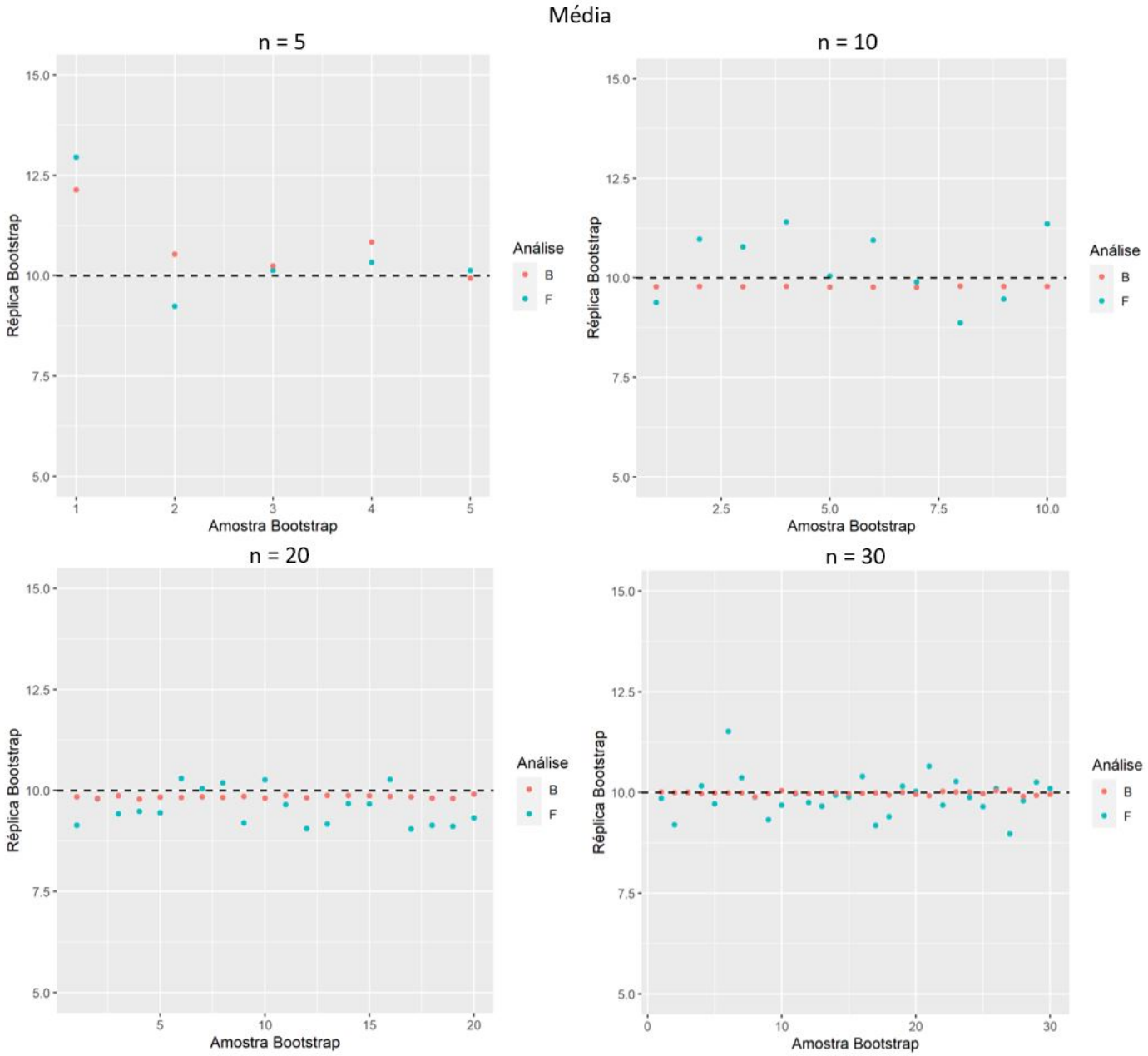

Figura 4.11. Estimativas *bootstrap* para média utilizando *priori* informativa

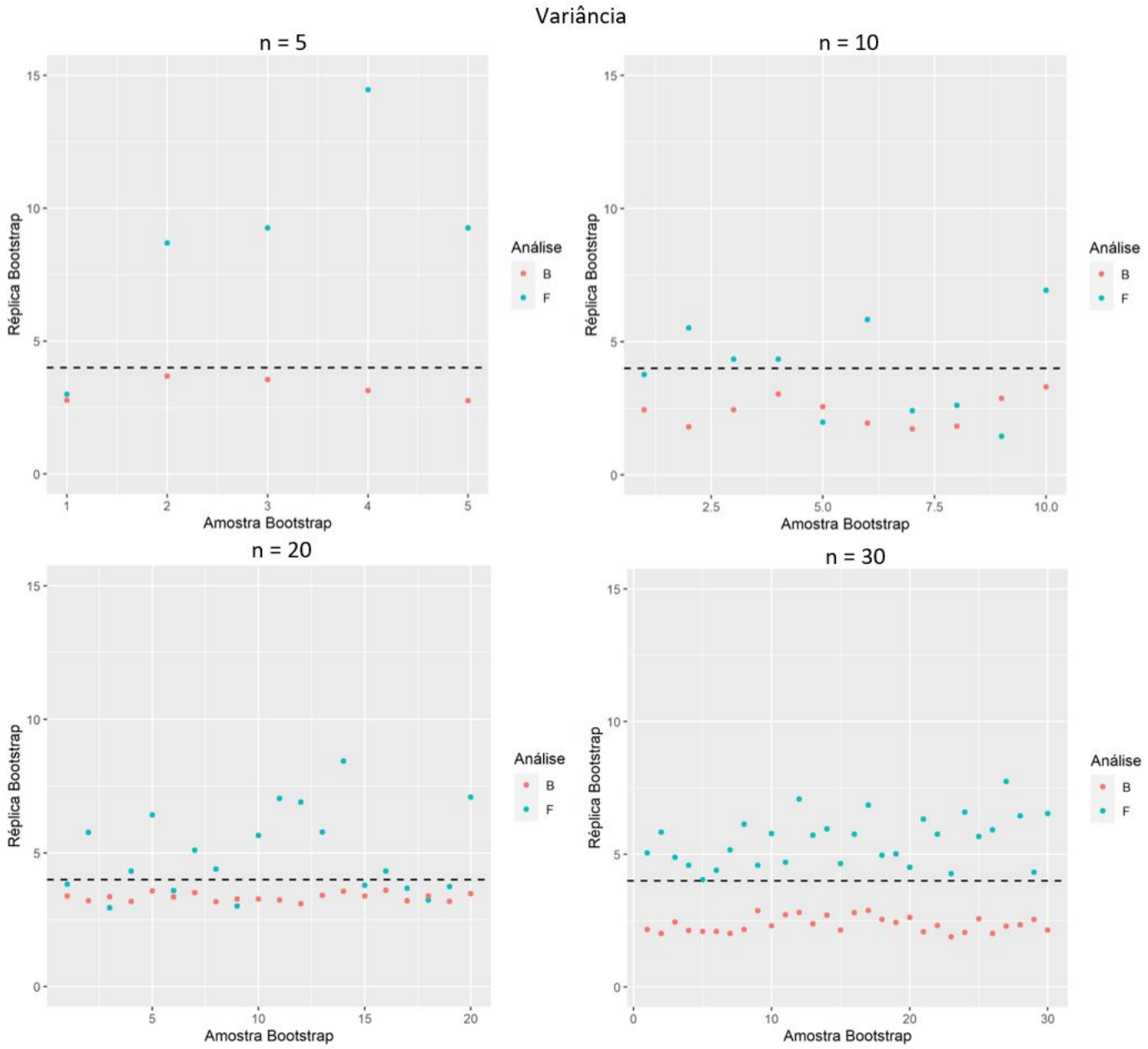

Figura 4.12. Estimativas *bootstrap* para variância utilizando *priori* informativa

As figuras 4.1 a 4.12 apresentam-se na forma de painel, tendo cada um deles fixado o parâmetro a ser estimado, o método de reamostragem utilizado e o tipo de *priori*. Cada painel é formado por 4 gráficos, que representam os 4 diferentes números de elementos por amostra (valores de n). A linha preta tracejada é o valor real do parâmetro a ser utilizado como referência, os pontos vermelhos e azuis são as réplicas (*jackknife* ou *bootstrap*) para o parâmetro utilizando análise bayesiana (B) e frequentista (F), respectivamente.

Ao observar as réplicas (representadas pelos pontos) isoladamente, os métodos de reamostragem *jackknife* e *bootstrap* assemelharam-se bastante, proporcionando uma mesma interpretação. O que mais interferiu, de fato, nos resultados das análises foi o número de elementos da amostra e a qualidade da *priori* considerada.

Conforme o esperado, é nítido que à medida que o tamanho amostral aumenta, o desempenho de ambas as estimações frequentista e bayesiana melhoram, fato representado pelos pontos vermelhos e azuis aproximando-se cada vez mais da linha preta tracejada, que simboliza o valor de referência. Aparentemente, a performance do método bayesiano é melhor na estimação da média, enquanto que o método frequentista é superior na estimação da variância.

Outro ponto importante é que ao utilizar uma *priori* informativa, o desempenho do método bayesiano é melhor até mesmo em amostras com apenas 5 elementos, o que não se fez regra quando considerou-se uma *priori* pouco informativa, corroborando o fato de que o método bayesiano depende fortemente de uma *priori* relevante para que possa, efetivamente, apresentar resultados confiáveis mesmo com amostras pequenas.

4.2 Dados Reais

4.2.1 Análise exploratória

As Figuras 4.13 até 4.17 representam *boxplots* para os dados de *Licania humilis* separados por localidade, sendo cada gráfico correspondente a um atributo foliar, especificados no título das imagens seguidos por suas unidades de medida. As letras C, P e R na legenda das figuras correspondem aos três diferentes ambientes considerados nesse estudo, sendo eles o Cerrado natural, sub-bosque de *Pinus* e Cerrado em regeneração, respectivamente.

Os *boxplots*, também conhecidos como gráficos de caixa, são uma ótima ferramenta para observar a dispersão dos dados, uma vez que deixam transparecer o valor da mediana, do primeiro e terceiro quartis, a distância interquartílica, além da presença de *outliers* (dados discrepantes, representados no gráfico por pontos isolados fora do corpo retangular do *boxplot*). Através deles também é possível observar se a dispersão dos dados é aproximadamente simétrica e se sua distribuição possui um formato aproximadamente normal.

Os gráficos nos permitem observar que a mudança de localidade parece afetar cada atributo da planta de maneira diferente. Esse fato é coerente, uma vez que cada estrutura possui uma função distinta nos processos fisiológicos da planta. Para avaliar se essa diferença é estatisticamente relevante ou não, foi necessário realizar uma Análise da Variância para cada variável resposta, em que os tratamentos são as diferentes localidades.

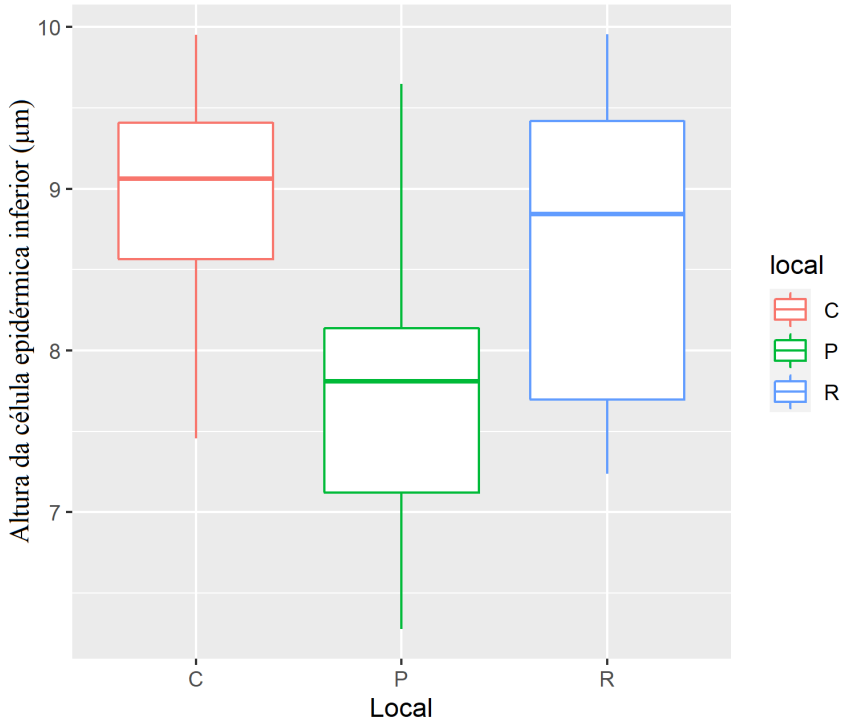

Figura 4.13. Altura da célula epidérmica inferior (μm) nas três localidades

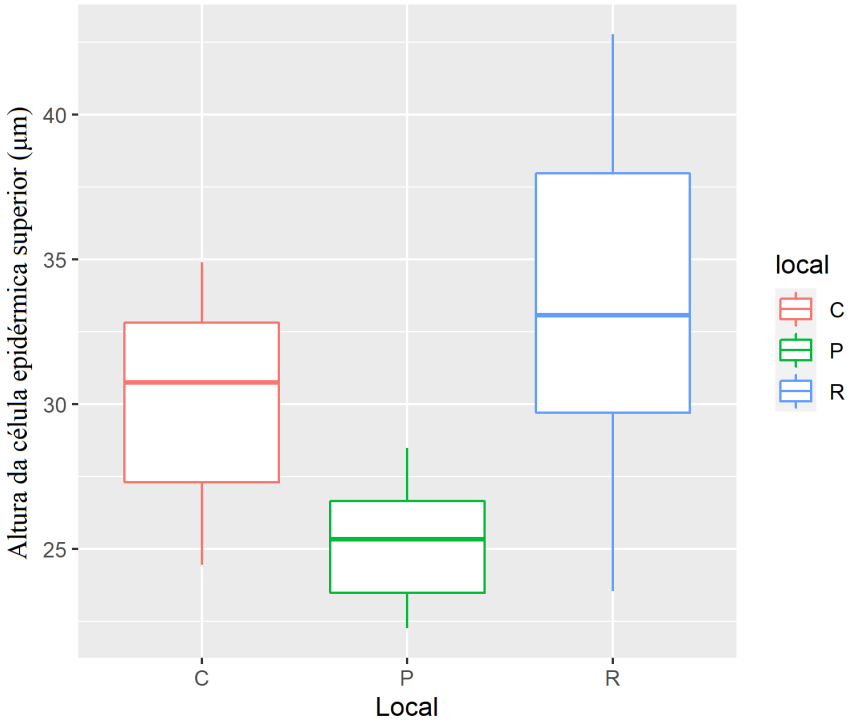

Figura 4.14. Altura da célula epidérmica superior (μm) nas três localidades

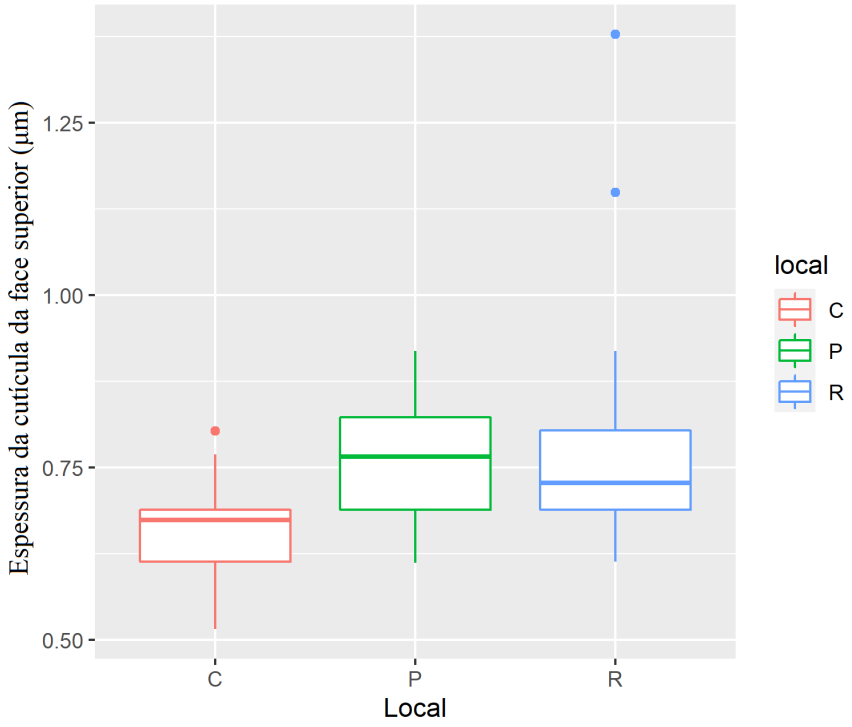

Figura 4.15. Espessura da cutícula da face superior (μm) nas três localidades

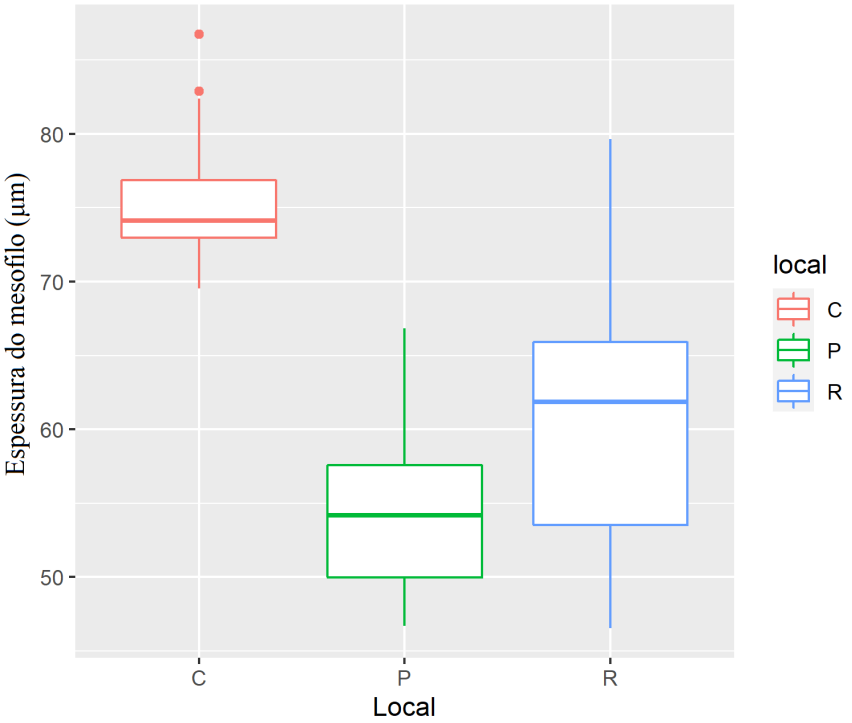

Figura 4.16. Espessura do mesofilo (μm) nas três localidades

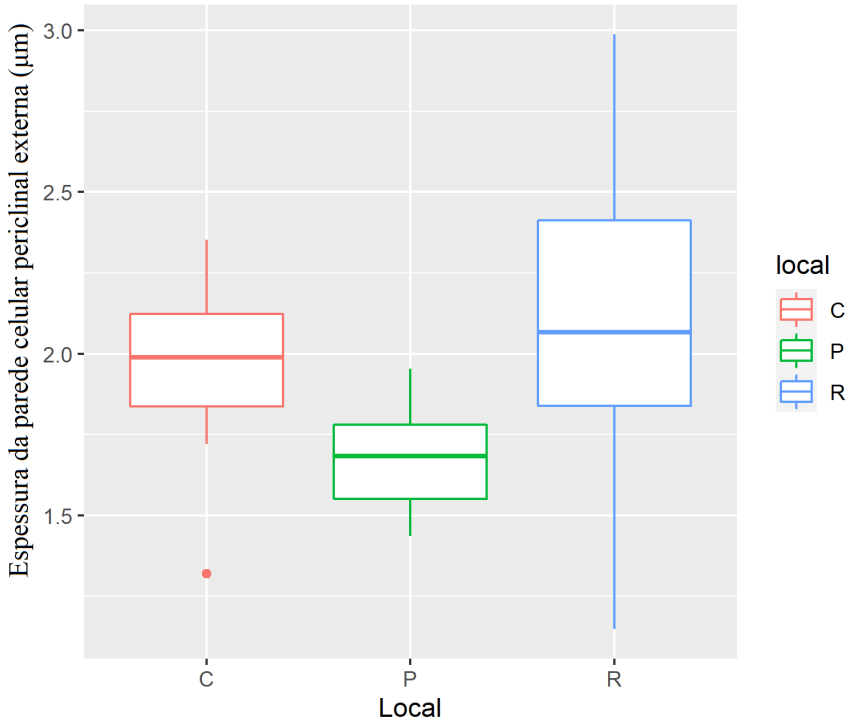

Figura 4.17. Espessura da parede celular periclinal externa (μm) nas três localidades

4.2.2 Análise frequentista

Embora o estudo de simulação realizado previamente tenha sugerido que a abordagem bayesiana possui uma melhor performance quando utilizadas amostras de tamanho pequeno, os dados reais de *L. humilis* foram analisados utilizando ambos os métodos frequentista e bayesiano, a fim de realizar uma comparação mais completa dos procedimentos e verificar os métodos de estimação também influenciam no resultado dos testes de comparação de médias.

As tabelas 4.5 a 4.9 contêm os resultados das Análises da Variância que foram realizadas com o objetivo de verificar se há diferença entre as plantas desenvolvidas no Cerrado natural, Cerrado em regeneração e sub-bosque de *Pinus* quanto à medida de 5 atributos anatômicos foliares. As estimativas das médias de cada ambiente foram obtidas através da estimação pelo método clássico, ou seja, utilizando-se a própria média amostral calculada sobre cada localidade.

As abreviações GL, SQ, QM e F apresentadas nas tabelas correspondem, respectivamente, ao grau de liberdade, soma de quadrados, quadrado médio e valor do teste F. A interpretação dese tipo de análise é dada pela observação da última coluna da tabela, a qual contém o P valor. Se este número for seguido por três asteriscos (***), significa que a diferença é significativa a 1%. Dois asteriscos (**) implicam em diferença significativa a 5%, e um asterisco (*), significativa a 10%. Considerando-se o nível mínimo de significância aceito para esse estudo como sendo 5%, todos os atributos anatômicos foliares avaliados mostraram-se estatisticamente diferentes conforme o local em que as plantas estavam localizadas.

Tabela 4.5. ANOVA para altura da célula epidérmica inferior (μm) considerando médias obtidas por meio de inferência frequentista

	GL	SQ	QM	F	P valor
Local	2	13,499	6,7495	10,088	0,0001611 ***
Resíduo	62	41,481	0,6691		

Tabela 4.6. ANOVA para altura da célula epidérmica superior (μm) considerando médias obtidas por meio de inferência frequentista

	GL	SQ	QM	F	P valor
Local	2	584,59	292,295	20,479	$1,485.10^{-7}$ ***
Resíduo	62	884,91	14,273		

Tabela 4.7. ANOVA para espessura da cutícula da face superior (μm) considerando médias obtidas por meio de inferência frequentista

	GL	SQ	QM	F	P valor
Local	2	0,20946	0,104728	6,1447	0,003676 **
Resíduo	62	1,05670	0,017044		

Tabela 4.8. ANOVA para espessura do mesofilo (μm) considerando médias obtidas por meio de inferência frequentista

	GL	SQ	QM	F	P valor
Local	2	4979,7	2489,86	50,824	$8,568.10^{-14}$ ***
Resíduo	62	3037,4	48,99		

Tabela 4.9. ANOVA para espessura da parede celular periclinal externa (μm) considerando médias obtidas por meio de inferência frequentista

	GL	SQ	QM	F	P valor
Local	2	1,7938	0,8969	9,1989	0,0003174 ***
Resíduo	62	6,0450	0,0975		

Uma vez concluído que pelo menos duas médias de localidade diferem entre si para todas as variáveis, foi necessário realizar um teste de Tukey para cada atributo foliar. A Tabela 4.10 apresenta os resultados desse testes, sendo cada coluna correspondente a um atributo, e cada linha correspondente a uma localidade. Nas colunas, médias estatisticamente diferentes estão representadas por letras diferentes, sendo a letra a o maior valor possível.

Com base nos resultados da análise com estimação frequentista, foi possível tirar as seguintes coclusões: Plantas que desenvolveram-se no sub-bosque de *Pinus* apresentaram valores menores que nos Cerrados natural e em regeneração para todos os atributos, exceto a espessura da cutícula da face superior, para a qual não apresentou diferença significativa de nenhum dos demais ambientes. Tanto para a altura da célula epidérmica superior como para a espessura da cutícula, os indivíduos situados no Cerrado em regeneração apresentaram valores maiores que aqueles no Cerrado natural, e o inverso ocorreu em relação à espessura do mesofilo. Quanto à altura da célula epidérmica inferior e a espessura da parede celular periclinal externa, plantas do Cerrado natural e do Cerrado em regeneração não apresentaram diferença estatística entre si.

Tabela 4.10. Teste de Tukey a 5% para todos os atributos foliares, considerando médias obtidas por meio de inferência frequentista

	C. E. Inferior	C. E. Superior	Cutícula	Mesofilo	P. Periclinal
Cerrado em regeneração	8,642 a	33,121 a	0,800 a	60,359 b	2,102 a
Cerrado natural	8,955 a	30,052 b	0,672 b	75,534 a	1,976 a
Sub-bosque de *Pinus*	7,767 b	25,226 c	0,757 ab	54,453 c	1,667 b

C.E. Inferior: Altura da célula epidérmica inferior (μm); C.E. Superior: Altura da célula epidérmica superior (μm); Cutícula: Espessura da cutícula da face superior (μm); Mesofilo: Espessura do mesofilo (μm); P. Periclinal: Espessura da parede celular periclinal externa (μm)

4.2.3 Análise bayesiana

As tabelas 4.11 a 4.15 apresentam os resultados das Análises da Variância que têm por finalidade avaliar se as médias das três localidades estudadas diferem significativamente entre si considerando os 5 atributos anatômicos foliares já descritos. Desta vez, os valores utilizados como as médias de cada ambiente foram as médias das distribuições *posteriori* estimadas através da inferência bayesiana para cada característica em cada tipo de vegetação.

A interpretação das tabelas se dá da mesma forma que na análise frequentista, observando o P valor e o número de asteriscos que o seguem. Vale ressaltar que os altos valores para o grau de liberdade do resíduo nessas tabelas se devem ao fato de que, na análise bayesiana, não são utilizados os dados da amostra propriamente dita, mas os dados da distribuição *posteriori* obtidos pelo pacote *stan*, que tendem a ser numerosos para que a os dados da *posteriori* apresentem características bem definidas da distribuição de probabilidade que ela representa.

Para todas as cinco variáveis resposta consideradas, o P valor foi marcado com três asteriscos (***), indicando que é inferior a 0,01 e, portanto, rejeitando-se a hipótese de nulidade correspondente ao nível 1%. Isso indica que, para todos os atributos do estudo, pelo menos duas médias de localidade diferem entre si, sugerindo que a espécie apresenta plasticidade anatômica foliar para todas as cinco características avaliadas.

Tabela 4.11. ANOVA para altura da célula epidérmica inferior (μm) considerando médias obtidas por meio de inferência bayesiana

	GL	SQ	QM	F	P valor
Local	2	63076	31538,1	637,22	$< 2,2.10^{-16}$ ***
Resíduo	5997	296811	49,5		

Tabela 4.12. ANOVA para altura da célula epidérmica superior (μm) considerando médias obtidas por meio de inferência bayesiana

	GL	SQ	QM	F	P valor
Local	2	35515	17757,3	285,31	$< 2,2.10^{-16}$ ***
Resíduo	5997	373252	62,2		

Em seguida, foi realizado um Teste de Tukey a 5% de significância, a fim de discriminar quais das médias de ambientes diferem significativamente entre si. Os referidos resultados estão apresentados na Tabela 4.16. Vale ressaltar que os valores mais altos de médias obtidas através desse método se devem à influência exercida pelos dados da espécie *Duguetia furfuracea*, utilizada como *priori*.

Tabela 4.13. ANOVA para espessura da cutícula da face superior (μm) considerando médias obtidas por meio de inferência bayesiana

	GL	SQ	QM	F	P valor
Local	2	27,896	13,9481	703,1	$< 2,2.10^{-16}$ ***
Resíduo	5997	118,968	0,0198		

Tabela 4.14. ANOVA para espessura do mesofilo (μm) considerando médias obtidas por meio de inferência bayesiana

	GL	SQ	QM	F	P valor
Local	2	104052991	52026495	1334,9	$< 2,2.10^{-16}$ ***
Resíduo	5997	233728872	38974		

Tabela 4.15. ANOVA para espessura da parede celular periclinal externa (μm) considerando médias obtidas por meio de inferência bayesiana

	GL	SQ	QM	F	P valor
Local	2	153,119	76,560	14745	$< 2,2.10^{-16}$ ***
Resíduo	5997	31,138	0,005		

Tabela 4.16. Teste de Tukey a 5% para todos os atributos, considerando médias obtidas por meio de inferência bayesiana

	C. E. Inferior	C. E. Superior	Cutícula	Mesofilo	P. Periclinal
Cerrado em regeneração	16,284 a	33,277 a	0,887 b	203,274 b	2,100 a
Cerrado natural	9,358 b	30,401 b	0,751 c	397,962 a	1,979 b
Sub-bosque de *Pinus*	9,455 b	27,319 c	0,902 a	77,879 c	1,718 c

C.E. Inferior: Altura da célula epidérmica inferior (μm); C.E. Superior: Altura da célula epidérmica superior (μm); Cutícula: Espessura da cutícula da face superior (μm); Mesofilo: Espessura do mesofilo (μm); P. Periclinal: Espessura da parede celular periclinal externa (μm)

A altura da célula epidérmica inferior é maior nas plantas do Cerrado em regeneração, mas não difere entre Cerrado natural e sub-bosque de *Pinus*. A altura da célula epidérmica superior e a espessura da parede celular periclinal externa apresentam o mesmo padrão de comportamento: os indivíduos do Cerrado em Regeneração apresentam valores maiores que os do Cerrado natural, que por sua vez superam aqueles do sub-bosque de *Pinus*. A cutícula foliar da face superior é mais espessa nas plantas do sub-bosque de *Pinus*, e menos espessa na Cerrado natural. Por fim, a espessura do mesofilo apresenta seu maior valor no Cerrado natural, e menor valor no sub-bosque de *Pinus*.

Fica evidente que cada tipo de ambiente está associado a diferentes exigências no que se refere à expressão dos atributos anatômicos foliares estudados. Utilizando-se como exemplo a cutícula foliar, é sabido que sua função na planta está vinculada à proteção contra perda de água. Dessa forma, justifica-se o fato de que a espessura da cutícula cresce conforme o aumento da hostilidade externa, sendo mais estreita no Cerrado natural (ambiente para o qual a espécie já é bem adaptada), e mais espessa no sub-bosque de *Pinus* (agressiva competição com espécie exótica).

4.3 Regressão

A Figura 4.18 apresenta um *heatmap* para os atributos foliares dos dados disponíveis. O objetivo do gráfico é determinar quais as variáveis mais correlacionada entre si, sendo que quanto mais

escura a cor, maior a correlação. A partir desse cálculo, é possível observar que os dois atributos mais fortemente correlacionados são altura do mesofilo e altura da célula epidérmica inferior, sendo definidas, respectivamente, como variável resposta e variável preditora.

As análises da variância para definição do grau do polinômio apontaram para um polinômio do primeiro grau como melhor opção, sugerindo que existe uma dependência linear entre as duas variáveis. A partir dessa informação, fez-se necessário obter os dois coeficientes dessa função, levando em consideração os princípios da validação cruzada. Assim, cada *training set* resultou em seus próprios coeficientes, evitando um possível sobreajuste decorrente da realização de uma única regressão linear utilizando a totalidade dos dados.

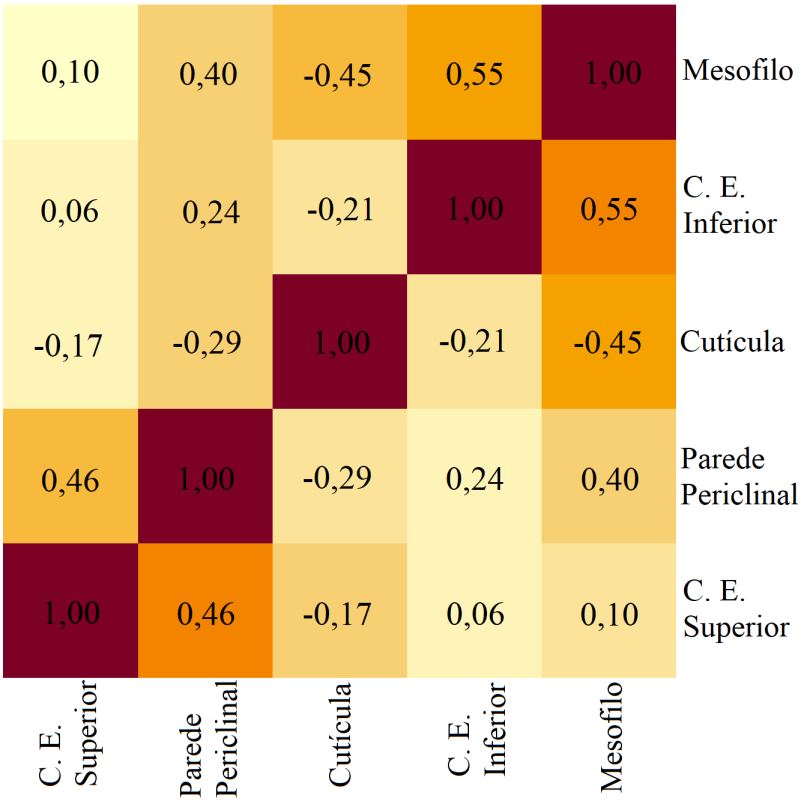

Figura 4.18. *Heatmap* para os atributos foliares

A Figura 4.19 apresenta um comparativo entre os valores reais de altura do mesofilo de indivíduos de *L. humilis* (pontos azuis) e os valores preditos deste atributo através de regressão linear (pontos vermelhos) utilizando valores da altura da célula epidérmica inferior como variável preditora.

É nítido que existe uma dispersão substancialmente maior nos dados reais em comparação aos preditos, o que é esperado quando se utiliza regressão polinomial. O fato é que a predição, nesse caso moldada por uma função do primeiro grau, busca acomodar os dados próximos a uma linha reta, não sendo capaz de capturar o efeito do acaso presente nos dados originais.

As Tabelas 4.17 e 4.18 apresentam os resultados das análises da variância para altura do mesofilo variando conforme os três diferentes ambientes. A diferença é que a Tabela 4.17 utiliza os dados reais da variável, enquanto que a 4.18 utiliza os valores preditos. Em ambos os casos, o P-valor foi significativo

Figura 4.19. Valores preditos *vs* valores reais

a 1%, ou seja, em pelo menos dois ambientes os indivíduos diferem significativamente entre si quanto à altura do mesofilo. Isso é um indício de boa performance da predição, já que esta foi capaz de manifestar corretamente a informação presente nos dados originais.

Tabela 4.17. ANOVA para dados reais de Altura do Mesofilo (μm) utilizando as médias reais

	GL	SQ	QM	F	P valor
Local	2	4979,9	2489,86	50,824	$8,568.10^{-14}$ ***
Resíduo	62	3037,4	48,99		

Tabela 4.18. ANOVA para dados preditos de Altura do Mesofilo (μm) utilizando as médias preditas

	GL	SQ	QM	F	P valor
Local	2	647,13	323,56	9,5812	0,0002367 ***
Resíduo	62	2093,78	33,77		

A Tabela 4.19 apresenta os testes de Tukey para comparar as médias de ambientes quanto à altura do mesofilo, discriminando os dados reais e preditos em diferentes colunas. Em ambos os casos, indivíduos no sub-bosque de *Pinus* diferiram dos outros dois ambientes, apresentando sempre a menor média. Entretando, nos dados reais os indivíduos no Cerrado natural e no Cerrado em regeneração diferiram entre si, o que não ocorreu para os dados preditos.

Tabela 4.19. Teste de Tukey para Altura do Mesofilo

	Dados reais	Dados preditos
Cerrado natural	75,53 a	67,39 a
Cerrado em regeneração	60,36 b	65,61 a
Sub-bosque de Pinus	54,45 c	59,24 b

Visto isso, a regressão linear mostrou-se capaz de captar as principais tendências da expressão de uma variável com base na informação de outra, mas não foi tão sutil de forma a transparecer toda a disparidade presente nos dados originais. É previsível que um teste de comparação de médias acuse maior diferença para dados reais do que preditos, devido à questão da dispersão dos dados apontada anteriormente.

5 CONCLUSÃO

Foi possível concluir, a partir dos resultados das análises, que os três ambientes estudados (Cerrado natural, Cerrado em regeneração e sub-bosque de *Pinus*) exercem efeitos diferentes sobre a expressão anatômica e micromorfológica foliar de *Licania humilis*. Essa diferença pode ser interpretada como uma estratégia adaptativa da espécie, de forma a ampliar suas chances de sobrevivência em cada um dos ambientes e suas respectivas condições externas.

Uma vez que a espécie vegetal em estudo é nativa do Cerrado, fica claro o motivo de os indivíduos nos outros ambientes manifestarem características mais distintas. Quando comparado com o Cerrado natural, a incidência luminosa no Cerrado em regeneração é maior, enquanto que no sub-bosque de *Pinus* predomina o sombreamento. Analisando altura do mesofilo, confirma-se que as folhas mais expostas ao sol apresentam uma maior espessura, enquanto que as folhas de sombra são mais delgadas.

Quanto à competição por água e nutrientes, faz-se muito mais agressiva no sub-bosque de *Pinus*, pois além da alta densidade de plantio quando comparada à vegetação típica savânica, envolve espécies exóticas que não fazem parte da dinâmica natural daquele ecossistema. No Cerrado em regeneração, a competição é ainda menor que no Cerrado natural, já que um menor número de plantas está presente. Nota-se que a espessura da cutícula, estrutura especializada na proteção das folhas, cresce conforme o ambiente vai tornando-se menos favorável para a espécie, atingindo sua maior espessura no sub-bosque de *Pinus*. Tudo isso sinaliza que *L. humilis* apresenta uma marcante plasticidade fenotípica quando exposta em diferentes condições externas.

Quanto aos métodos de inferência aplicados a amostras com poucos elementos, a abordagem bayesiana destacou-se no estudo de simulação, além de apresentar resultados coerentes na análise com dados reais. É compreensível que a performance do método tenha superado a da abordagem clássica nesse caso, já que a *priori* complementa a informação da amostra que, a princípio, é escassa.

Em relação à regressão polinomial para predição de dados, os resultados das análises foram coerentes, mas obviamente há uma perda de informação em relação ao que seria se fosse possível medir a variável de fato. Ainda assim, é uma ferramenta interessante à qual é possível recorrer como alternativa quando a variável de interesse for realmente de difícil mensuração, mas jamais deve ser utilizada para substituir os dados originais quando se pode ter acesso a eles.

REFERÊNCIAS

Acharya, A. S., A. Prakash, P. Saxena, and A. Nigam, 2013 Sampling: Why and how of it. Indian Journal of Medical Specialties **4**: 330–333.

Appezzato-da Glória, B. and S. M. Carmello-Guerreiro, 1992 *Anatomia vegetal*. Universidade de São Paulo. ESALQ.

Banner, K. M., K. M. Irvine, and T. J. Rodhouse, 2020 The use of bayesian priors in ecology: The good, the bad and the not great. Methods in Ecology and Evolution **11**: 882–889.

Bernardo, J. M., M. Bayarri, J. O. Berger, A. Dawid, D. Heckerman, A. F. Smith, and M. West, 2011 *Bayesian statistics 9*. Oxford University Press.

Bernardo, J. M. and A. F. Smith, 2009 *Bayesian theory*, volume 405. John Wiley & Sons.

Berrar, D., 2019 Cross-validation. Tokyo Institute of Technology.

Bieras, A. C. and M. d. G. Sajo, 2009 Leaf structure of the cerrado (brazilian savanna) woody plants. Trees **23**: 451–471.

Bolfarine, H. and M. C. Sandoval, 2001 *Introdução à inferência estatística*, volume 2. SBM.

Bradshaw, A., 1965 Evolutionary significance of phenotypic plasticity in plants. volume 13 of *Advances in Genetics*, pp. 115–155, Academic Press.

Butler, R., 2015 Destructive sampling ethics. Nature Geoscience **8**: 817–818.

Cam, L. L., 1990 Maximum likelihood: An introduction. International Statistical Review / Revue Internationale de Statistique **58**: 153–171.

Casella, G. and R. L. Berger, 2021 *Statistical inference*. Cengage Learning.

Chernick, M. R., 2012 Resampling methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **2**: 255–262.

Cutler, D. F., T. Botha, and D. W. Stevenson, 2009 *Anatomia vegetal: uma abordagem aplicada*. Artmed Editora.

da Silva, F. A. M., E. D. Assad, E. T. Steinke, and A. G. Müller, 2008 Clima do bioma cerrado. Agricultura tropical: quatro décadas de inovações tecnológicas, institucionais e políticas. ALBUQUERQUE, ACS pp. 93–148.

Dantas, C. A. B., 2013 *Probabilidade: Um Curso Introdutório Vol. 10*. Edusp.

de Mendiburu, F., 2021 agricolae tutorial (version 1.3-5). Universidad Nacional Agraria: La Molina, Peru .

Dietterich, T., 1995 Overfitting and undercomputing in machine learning. ACM computing surveys (CSUR) **27**: 326–327.

Diggle, P. J., A. G. Chetwynd, and A. Chetwynd, 2011 *Statistics and scientific method: An introduction for students and researchers*. Oxford University Press.

Dixon, P. M., 2006 Bootstrap resampling. Encyclopedia of environmetrics .

Efron, B., 1992 Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pp. 569–593, Springer.

Efron, B. and R. J. Tibshirani, 1994 *An introduction to the bootstrap*. CRC press.

Fank-de Carvalho, S., M. Gomes, P. Tanno, and S. Báo, 2010 Leaf surfaces of gomphrena spp. (amaranthaceae) from cerrado biome. Biocell : official journal of the Sociedades Latinoamericanas de Microscopía Electronica ... et. al **34**: 23–35.

Fisher, R. A., 1935/1960 The design of experiments (7th ed.). New York: Hafner Pub .

Fox, G. A., S. Negrete-Yankelevich, and V. J. Sosa, 2015 *Ecological statistics: contemporary theory and application*. Oxford University Press, USA.

Glickman, M. E. and D. A. Van Dyk, 2007 Basic bayesian methods. Topics in Biostatistics pp. 319–338.

Good, P., 2013 *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

Gotelli, N., 2009 Ecologia. editora planta. Londrina, PR .

Griffiths, A., S. Wessler, S. Carroll, and J. Doebley, 2010 *Introduction to Genetic Analysis*. W. H. Freeman.

Hairston, N. G., 1989 *Ecological experiments: purpose, design and execution*. Cambridge University Press.

Hanson, H. C., 1917 Leaf-structure as related to environment. American Journal of Botany pp. 533–560.

Hooton, J. W., 1991 Randomization tests: statistics for experimenters. Computer Methods and Programs in Biomedicine **35**: 43–51.

Ivanova, L. and V. P'yankov, 2002 Structural adaptation of the leaf mesophyll to shading. Russian Journal of Plant Physiology **49**: 419–431.

James, G., D. Witten, T. Hastie, and R. Tibshirani, 2013 *Resampling Methods*. Springer New York, New York, NY.

Javelle, M., V. Vernoud, P. M. Rogowsky, and G. C. Ingram, 2011 Epidermis: the formation and functions of a fundamental plant tissue. New Phytologist **189**: 17–39.

Johnson, R. W., 2001 An introduction to the bootstrap. Teaching statistics **23**: 49–54.

Klink, C. A. and R. B. Machado, 2005 A conservação do cerrado brasileiro. Megadiversidade **1**: 147–155.

Krus, D. J. and E. A. Fuller, 1982 Computer assisted multicrossvalidation in regression analysis. Educational and Psychological Measurement **42**: 187–193.

Kurtz, A. K., 1948 A research test of the rorschach test. Personnel Psychology .

Kwak, S. G. and J. H. Kim, 2017 Central limit theorem: the cornerstone of modern statistics. Korean journal of anesthesiology **70**: 144–156.

Lemoine, N. P., 2019 Moving beyond noninformative priors: why and how to choose weakly informative priors in bayesian analyses. Oikos **128**: 912–928.

Lenk, P. and B. Orme, 2009 The value of informative priors in bayesian inference with sparse data. Journal of Marketing Research **46**: 832–845.

Liang, K.-Y. and S. L. Zeger, 1993 Regression analysis for correlated data. Annual review of public health **14**: 43–68.

Lindley, D. V., 1965 *Introduction to probability and statistics: from a Bayesian viewpoint. 2. Inference*. CUP Archive.

Mackay, T. F. C., E. A. Stone, and J. F. Ayroles, 2009 The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics **10**.

Magalhães, M. N., 2006 *Probabilidade e variáveis aleatórias*. Edusp.

Marques, A., Q. Garcia, J. Rezende, and G. Fernandes, 2000 Variations in leaf characteristics of two species of miconia in the brazilian cerrado under different light intensities. Tropic. Ecol. **41**.

Mosier, C. I., 1951 Problems and designs of cross-validation. Educational and Psychological Measurement **11**: 5–11.

Murteira, B. J. F., 1995 Introdução à inferência bayesiana ISSN 0872-895X.

Ostertagová, E., 2012 Modelling using polynomial regression. Procedia Engineering **48**: 500–506.

Poehlman, J. M., 2013 *Breeding field crops*. Springer Science & Business Media.

Pompelli, M. F., K. R. Mendes, M. V. Ramos, J. N. Santos, D. T. Youssef, J. D. Pereira, L. Endres, A. Jarma-Orozco, R. Solano-Gomes, B. Jarma-Arroyo, *et al.*, 2019 Mesophyll thickness and sclerophylly among calotropis procera morphotypes reveal water-saved adaptation to environments. Journal of Arid Land **11**: 795–810.

Prajapati, B., M. Dunne, and R. Armstrong, 2010 Sample size estimation and statistical power analyses. Optometry today **16**: 10–18.

Quenouille, M. H., 1949 Approximate tests of correlation in time-series. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 45, pp. 483–484, Cambridge University Press.

R Core Team, 2017 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Reis, E., P. Melo, R. Andrade, and T. Calapez, 1999 Estatística aplicada. Lisboa: Edições Sílabo pp. 21–26.

Resetarits, W. J. and J. Bernardo, 2001 *Experimental ecology: issues and perspectives*. Oxford University Press on Demand.

Ristic, Z. and M. A. Jenks, 2002 Leaf cuticle and water loss in maize lines differing in dehydration avoidance. Journal of Plant Physiology **159**: 645–651.

Rôças, G., C. F. Barros, and F. R. Scarano, 1997 Leaf anatomy plasticity of alchornea triplinervia (euphorbiaceae) under distinct light regimes in a brazilian montane atlantic rain forest. Trees **11**: 469–473.

Rodrigues, W. C. *et al.*, 2007 Metodologia científica. Faetec/IST. Paracambi pp. 2–20.

Rohatgi, V. K., 2013 *Statistical inference*. Courier Corporation.

Rosenbaum, P. R., 2005 Observational study. Encyclopedia of statistics in behavioral science .

Rossatto, D. R. and R. M. Kolb, 2010 Gochnatia polymorpha (less.) cabrera (asteraceae) changes in leaf structure due to differences in light and edaphic conditions. Acta Botanica Brasilica **24**: 605–612.

Scheiner, S. M. and J. Gurevitch, 2001 *Design and analysis of ecological experiments*. Oxford University Press.

Shao, J. and D. Tu, 2012 *The jackknife and bootstrap*. Springer Science & Business Media.

Shepherd, L. D., 2017 A non-destructive dna sampling technique for herbarium specimens. PloS one **12**: e0183555.

Silvey, S. D., 1975 *Statistical inference*, volume 7. CRC press.

Smid, S., D. McNeish, M. Miočević, and R. Van de Schoot, 2019 Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. Structural Equation Modeling: A Multidisciplinary Journal **27**: 1–31.

Stan Development Team, 2022 RStan: the R interface to Stan. R package version 2.21.7.

Taconeli, C. A. and S. R. Giolo, 2020 Maximum likelihood estimation based on ranked set sampling designs for two extensions of the lindley distribution with uncensored and right-censored data. Comput. Stat. **35**: 1827–1851.

Thompson, S. K., 2012 *Sampling*, volume 755. John Wiley & Sons.

Tukey, J., 1958 Bias and confidence in not quite large samples. Ann. Math. Statist. **29**: 614.

Van de Schoot, R. and M. Miocević, 2020 *Small sample size solutions: A guide for applied researchers and practitioners*. Taylor & Francis.

Via, S. and R. Lande, 1985 Genotype-environment interaction and the evolution of phenotypic plasticity. Evolution **39**: 505–522.

Wagenmakers, E.-J., M. Lee, T. Lodewyckx, and G. J. Iverson, 2008 Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses*, pp. 181–207, Springer.

Wang, X., Y. Xu, Z. Hu, and C. Xu, 2018 Genomic selection methods for crop improvement: Current status and prospects. The Crop Journal **6**: 330–340.

Yu, C. H., 2002 Resampling methods: concepts, applications, and justification. Practical Assessment, Research, and Evaluation **8**: 19.